analytical chemistry

Article

pubs.acs.org/ac

Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software

Hiroshi Tsugawa,[†] Tobias Kind,[‡] Ryo Nakabayashi,[†] Daichi Yukihira,[§] Wataru Tanaka,[∥] Tomas Cajka,[‡] Kazuki Saito,[†,⊥] Oliver Fiehn,*[,‡,#] and Masanori Arita*[,†,∥,○]

[†]RIKEN Center for Sustainable Resource Science, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

[‡]Genome Center, University of California—Davis, Davis, California 95616, United States

[§]Reifycs Inc., Minato-ku, Tokyo 105-0003, Japan

[∥]Department of Genetics, SOKENDAI (The Graduate University for Advanced Studies), 1111 Yata, Mishima, Shizuoka 411-8540, Japan

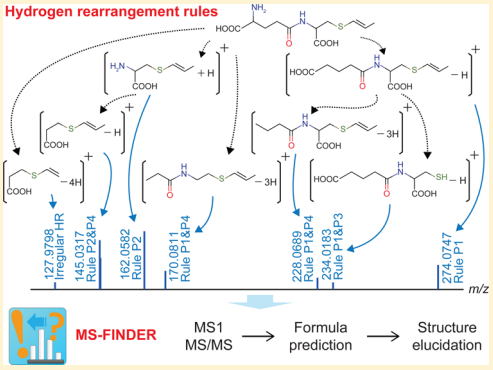
[⊥]Graduate School of Pharmaceutical Sciences, Chiba University, 1-33 Yayoi-cho, Inage-ku, Chiba 263-8522, Japan

[#]Biochemistry Department, King Abdulaziz University, Jeddah 21589, Saudi Arabia

[○]National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

Supporting Information

ABSTRACT: Compound identification from accurate mass MS/MS spectra is a bottleneck for untargeted metabolomics. In this study, we propose nine rules of hydrogen rearrangement (HR) during bond cleavages in low-energy collision-induced dissociation (CID). These rules are based on the classic even-electron rule and cover heteroatoms and multistage fragmentation. We evaluated our HR rules by the statistics of MassBank MS/MS spectra in addition to enthalpy calculations, yielding three levels of computational MS/MS annotation: "resolved" (regular HR behavior following HR rules), "semiresolved" (irregular HR behavior), and "formula-assigned" (lacking structure assignment). With this nomenclature, 78.4% of a total of 18506 MS/MS fragment ions in the MassBank database and 84.8% of a total of 36370 MS/MS fragment ions in the GNPS database were (semi-) resolved by predicted bond cleavages. We also introduce the MS-FINDER software for structure elucidation. Molecular formulas of precursor ions are determined from accurate mass, isotope ratio, and product ion information. All isomer structures of the predicted formula are retrieved from metabolome databases, and MS/MS fragmentations are predicted in silico. The structures are ranked by a combined weighting score considering bond dissociation energies, mass accuracies, fragment linkages, and, most importantly, nine HR rules. The program was validated by its ability to correctly calculate molecular formulas with 98.0% accuracy for 5063 MassBank MS/MS records and to yield the correct structural isomer with 82.1% accuracy within the top-3 candidates. In a test with 936 manually identified spectra from an untargeted HILIC-QTOF MS data set of human plasma, formulas were correctly predicted in 90.4% of the cases, and the correct isomer structure was retrieved at 80.4% probability within the top-3 candidates, including for compounds that were absent in mass spectral libraries. The MS-FINDER software is freely available at http://prime.psc.riken.jp/.



Untargeted metabolomics by liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) holds the promise for discovering new biochemical and physiological mechanisms, if structurally identified compounds can be quickly and correctly annotated.[1−3] There are four major criteria for metabolite identification: retention time, precursor *m/z*, isotopic ratio, and MS/MS spectra.[4] Of these, MS/MS spectra are the most informative for verifying molecular substructures and for distinguishing isomers. However, there are far fewer authentic mass spectra available in either commercial or free MS databases than chemical and metabolite structures compiled in metabolomic and chemical repositories. For example, MassBank, the public mass-spectral repository,[5] contains only 26296 MS/MS spectra covering 3127 authentic compound structures (disregarding stereochemistry), and only 9344 organic compounds are available from the commercial NIST MS/MS library (NIST14). In comparison, metabolome structure databases such as HMDB[6] (humans) and KNApSAcK[7] (plants) are much larger with 41993 and 50899 structures, respectively (downloaded on April 1, 2015). That means, even in an optimistic estimate, only about 5% of all known metabolites are MS/MS spectra available. For

Received: February 26, 2016
Accepted: July 15, 2016
Published: July 15, 2016

DOI: 10.1021/acs.analchem.6b00770
Anal. Chem. 2016, 88, 7946−7958

Table 1. Statistics of Hydrogen Rearrangements for CNOPS Elements in MassBank[a]

positive ion mode C 45486, N 7070, O 10789, S 664, P 128				negative ion mode C 20333, N 2278, O 9105, S 791, P 118			
cleavage	element	hydrogen rearrangement	count	cleavage	element	hydrogen rearrangement	count
first bond cleavage (ionized element behavior)	**C**	**0**	**2094**	first bond cleavage (ionized element behavior)	**C**	**0**	**247**
	C	*+1*	*160*		*C*	*−1*	*51*
	C	*+2*	*216*		**C**	**−2**	**418**
	N	*0*	*129*		**N**	**0**	**161**
	N	*+1*	*64*		*N*	*−1*	*16*
	N	**+2**	**581**		*N*	*−2*	*8*
	O	*0*	*23*		**O**	**0**	**357**
	O	*+1*	*26*		*O*	*−1*	*187*
	O	**+2**	**265**		*O*	*−2*	*4*
	*S	0	153		**S**	**0**	**28**
	S	*+1*	*5*		*S	−1	85
	S	**+2**	**18**		*S*	*−2*	*0*
	*P	0	29		**P**	**0**	**0**
	P	*+1*	*1*		*P*	*−1*	*2*
	P	**+2**	**0**		*P	−2	108
second and later bond cleavage (neutralized element behavior)	**C**	**−1**	**4245**	second and later bond cleavage (neutralized element behavior)	**C**	**+1**	**223**
	C	*0*	*145*		*C*	*0*	*41*
	C	**+1**	**538**		**C**	**−1**	**1102**
	N	**−1**	**276**		**N**	**+1**	**108**
	N	*0*	*58*		*N*	*0*	*13*
	N	**+1**	**878**		**N**	**−1**	**24**
	O	**−1**	**23**		**O**	**+1**	**119**
	O	*0*	*10*		*O*	*0*	*35*
	O	**+1**	**121**		**O**	**−1**	**64**
	S	**−1**	**19**		**S**	**+1**	**3**
	S	*0*	*5*		*S*	*0*	*10*
	S	**+1**	**11**		**S**	**−1**	**50**
	P	**−1**	**16**		**P**	**+1**	**0**
	P	*0*	*0*		*P*	*0*	*0*
	P	**+1**	**0**		**P**	**−1**	**26**

[a]A total 13260 (positive) and 5246 (negative) MS/MS fragment ions within 10-mDa mass accuracy were classified. Fragments matching the even-electron rule are bolded. Minor hydrogen rearrangement behaviors are shown in italic. An asterisk was added to the behaviors of P and S; they were rigorously incorporated in the rules of major hydrogen rearrangement behaviors. The first bond cleavage is the ionizing fragmentation, and the second and later bond cleavages are the fragmentation of ions, i.e., the hydrogens are rearranged to neutralize the structure.

comprehensively screening all known small molecules, there are over 60 million structures listed in the PubChem database.

For many reasons, it is not realistic to try measuring all natural compounds as authentic compounds to obtain their MS/MS spectra. The only alternative strategy is to computationally simulate the fragmentation process under low-energy collision-induced dissociation (CID) to obtain theoretical spectra of these input structures. Once a rigorous correspondence between the precursor structure and its diagnostic fragments in CID is established, the efficiency of identification in untargeted metabolomics will be markedly improved. A successful example for glycerolipids was presented in LipidBlast[8] and similar commercial libraries of in silico MS/MS spectra. Such theoretical libraries facilitated comprehensive lipid identification and led to the emergence of lipidomics.[4,9] However, the fragmentation of smaller metabolites, especially those containing heteroatoms, is not well understood. Many computational approaches have already approached the problem of precursor prediction from MS/MS spectra,[10−13] and machine-learning methods such as CFM-ID[12] and CSI:FingerID[13] are reported to reliably calculate chemical structures. In principle, these tools translate masses into molecular fragments through combinatorial structure generation. However, if we can rationalize the fragmentation process, especially the number of rearranged hydrogens, the objectivity and reliability of identification strategies will improve significantly.

We here present a strategy for computational MS/MS fragmentations with mechanistic insights in rearrangement of hydrogens during bond cleavages in low-energy CID-based fragmentation. A famous example is the cleavage of an N−C bond in the positive ion mode, where two hydrogens are attached to the nitrogen atom to represent the positive charge.[14,15] This observation was first formalized as the even-electron rule;[16,17] it states that even-electron cations rarely lose a radical to form an odd-electron cation. Based on extensively curated database records and enthalpy calculations, we here formulate "hydrogen rearrangement (HR) rules" by refining and extending the even-electron rule for carbon (C) and heteroatoms, oxygen (O), nitrogen (N), phosphorus (P), and sulfur (S).[18] We especially focus on the unique behaviors of P and S attributable to their higher electronic degree of freedom. We also discuss irregular HR behaviors as the exception to the classic even-electron rule based on our statistics.

We have implemented these HR rules into the MS-FINDER program for the elucidation of chemical structures from

DOI: 10.1021/acs.analchem.6b00770
Anal. Chem. 2016, 88, 7946−7958

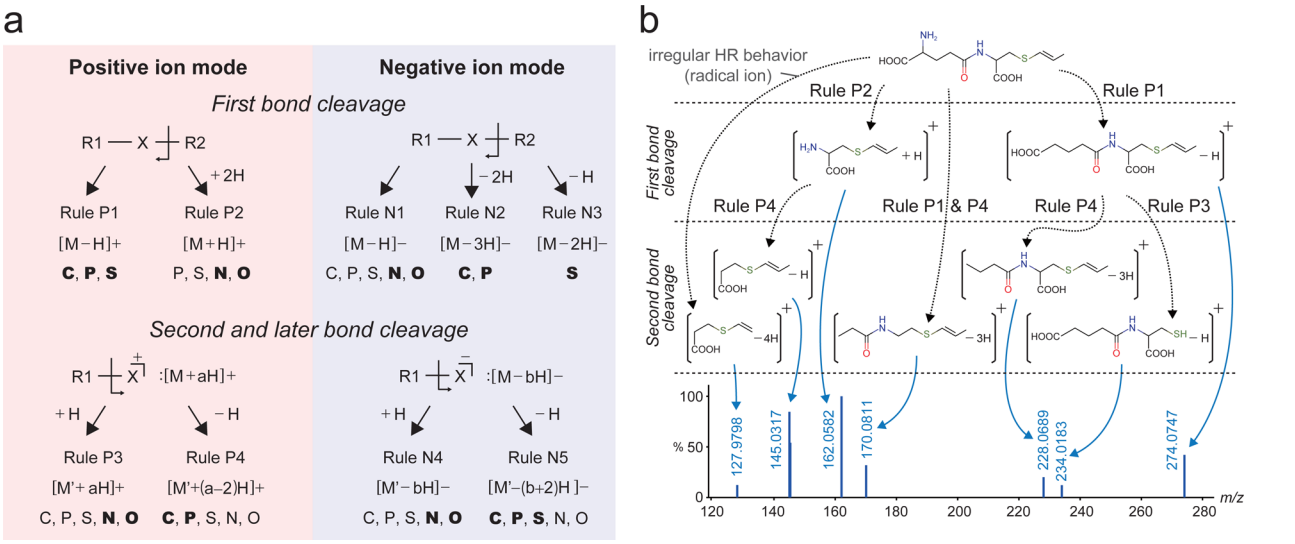


Figure 1. Rules of hydrogen rearrangement for characteristic fragments. (a) Rules P1, P2, N1, N2, and N3 are applied for the precursor structure described as R1-X-R2. The capital "X" indicates any one of the CNOPS elements. Candidate atoms are shown under each rule (bold face indicates the most frequent pattern). M and M′ indicate the neutralized (hydrogen-supplemented) form. The starting structure becomes R1-X for the second and later bond cleavages; the X part is assumed to be charged. * a, b: the total number of arranged hydrogens. (b) An example of in silico MS/MS annotation. Given a precursor structure (γ-glutamyl-*S*-1-carboxypropenylcysteine), the HR rules were combinatorially applied and obtained fragments were compared with the actually observed eight fragments. In this case, seven fragment ions were resolved and one ion (*m*/*z* 127.9798) was semiresolved as a radical ion.

accurate mass precursor ions and MS/MS spectra. Accuracy for correct formula predictions and the subsequent isomer annotation were assessed by querying 5063 spectra records from MassBank and by 936 spectra records of blood plasma from a human cohort study, including the elucidation of new compounds that were absent from reference mass spectral libraries. Finally, we introduce the comparison with other identification programs with three examples.

RESULTS AND DISCUSSION

Summary of Hydrogen Rearrangement Rules and Computational MS/MS Annotations. Bond cleavages in low-energy CID mostly produce fragments with an even number of electrons along with the addition or subtraction of protons. We call this process of electron/proton shifts collectively as hydrogen rearrangement (HR), even when no intramolecular rearrangement occurs. First, we examined the statistics of HR from curated MS/MS spectra of known precursors in the MassBank public repository (Table 1; also see Materials and Methods). We used high abundance MS/MS peaks that exceeded 10% of their base peaks to identify common features. In this study, the fragment ions are represented as a neutralized (i.e., valence-satisfied) structure plus or minus hydrogen(s) and shown as $[M \pm aH]^+$ or $[M \pm bH]^-$, where M stands for the neutralized structure, and a and b stand for the number of hydrogens. From Table 1 we can obtain three important observations: (1) even-electron ions, shown in boldface, predominate in CID-based fragmentation and their patterns can be formulated; (2) the rearrangement pattern of each heteroatom (N, O, P, S) differs; and (3) odd-electron ions and other anomalies are typically monitored at a frequency of about 10% (italic font).

Second, we formulated nine rules as regular HR behaviors in the cleavage process of C, N, O, P, and S (Figure 1a); there are four rules for the positive ion mode (rule P1−P4) and five rules for the negative mode (rule N1−N5). We decided to deal with irregular HR behaviors as the exception of the above rules for the MS/MS annotations, since the fragment ions from such behaviors can be stabilized by the formation of (hyper-) conjugation (see Treatment of Irregular HR Behaviors as Exception below). Consequently, we propose three levels for computational MS/MS annotation as follows: (1) mass peaks are considered resolved when they are explained by the nine HR rules; (2) mass peaks are semiresolved when they are explained within ±2 hydrogens, covering irregular HR behaviors. The semiresolved fragments are interpreted as the detection of odd-electron- or radical fragments, or as the result of a charge-transfer in even-electron fragments (see later sections); (3) For the remaining unresolved mass peaks, each one is assigned by formula and they may be interpreted as intramolecular rearrangement or as unknown isomerization. An example of γ-glutamyl-*S*-1-carboxypropenylcysteine is shown in Figure 1b, whose peak annotation requires all four rules of the positive ion mode in addition to one irregular HR behavior.

Third, we tested the performance of in silico MS/MS annotation by HR rules using a total of 13260 fragment ions listed in 3462 MS/MS spectra downloaded from MassBank (covering 1,157 unique compounds) in positive electrospray ionization (ESI), and 5246 fragment ions listed in 1601 MS/MS spectra (covering 818 unique compounds) in negative ESI mode. For ESI(+), 69.8% of the fragment ions were accurately resolved (70.1% for ESI(−)), 8.6% of the ESI(+) fragment ions were semiresolved (9.9% for ESI(−)), that is, resolved within ±2 hydrogens, and 18.8% of the ESI(+) fragment ions were assigned for their formulas only (16.7% for ESI(−)).

In order to ensure that there was no hidden structure bias by using MassBank metabolome spectra, we tested the generality of our computational MS/MS annotation using MS/MS records of natural products deposited in the GNPS database: 35738 fragment ions of 3712 spectra records for positive and 632 fragment ions of 147 spectra records for negative. While the registered molecule classes in GNPS were indeed

DOI: 10.1021/acs.analchem.6b00770
Anal. Chem. 2016, 88, 7946−7958

structurally different from MassBank, the result statistics closely resembled the MassBank result (Supporting Information, Table S1). In positive and negative ESI mode, 61.2% and 67.7% of the fragment ions were accurately resolved, 23.6% and 19.6% fragments were semiresolved, and 10.3% and 9.7% ions were assigned for their formulas. However, the unique behavior of sulfur, such as the preference of homolysis, was not evident in GNPS because of the scarcity of data, especially for ESI(−).

Hydrogen Rearrangement Rules for Understanding Characteristic Ions. In Figure 1a we summarize the regular HR behaviors and formulate the number of rearranged hydrogens for cleaved terminal C, N, O, P, and S fragment ions. These patterns are the superset of the classic even-electron rule and also cover multistage fragmentations. In low-energy CID, a rearrangement of C-bond cleavage adds no hydrogen (rule P1), whereas a rearrangement of N- or O-bond cleavages adds two hydrogens (rule P2). These rules correspond to the classic even-electron rule for C, N, and O. Behaviors of P and S are different: usually rule P1 (shown in bold),[19] and occasionally rule P2 applies. In fact, the statistics in Table 1 indicate that rule P2 is not observed in P-bond cleaved sites. However, manual inspection revealed that this observation was attributed to structural bias in our MS/MS data where almost all phosphorus appeared as phosphate. The important observation here is that the cleavage of P or S follows rule P1, whereas the cleavage of N or O follows rule P2.

We also formulated rule P3 and rule P4 for the second bond and subsequent cleavages for which starting ions are described as $[M + aH]^+$ ("a" stands for the number of hydrogens). We hypothesized that for all subsequent bond cleavages, fragments are singly charged. Consequently, the cleaved fragment is either neutralized by adding one hydrogen as $[M' + aH]^+$ (rule P3) or double-bonded (or cyclized) by losing one hydrogen as $[M' + (a - 2)H]^+$ (rule P4), where M′ stands for the substructure derived from a precursor M. Our statistics show that the loss of one hydrogen is more frequent for C and P while the addition of one hydrogen is more frequent for N and O (Table 1). As ring fission requires the breaking of two bonds, it is modeled as the combination of initial cleavage (rule P1 or P2) and the next cleavage (rule P3 or P4).

In ESI(−), fragment ions are stabilized after the initial bond cleavage without any hydrogen recruitment for all elements (CNOPS; rule N1). For C-bond cleavages, a two-hydrogen loss is more common, confirming a finding reported by Nakata[16] (rule N2). The same is true for the tetra-coordinated form of P, that is, phosphate. For S-bond cleavages, an odd-electron ion is the result of "homolysis" (rule N3), especially in sulfonates. Indeed, the abundance ratio of homolysis and heterolysis has been used as the diagnostic marker of glucosinolates (S-containing metabolites).[20] The ratio is affected by the molecular environment, especially the formation of coordinates.

For second and subsequent bond cleavages in ESI(−), the starting ion is described as $[M - bH]^-$ ("b" stands for the number of hydrogens). As in ESI(+), adding one hydrogen neutralizes the cleaved fragment; the notation of the fragment remains the same (rule N4). When one hydrogen is lost, the notation becomes $[M' - (b + 2)H]^-$ (rule N5). Our statistics indicate that the addition of one hydrogen is more frequent for N and O, while the loss of one hydrogen is more frequent for C, P, and S (Table 1). Thus, we regularized the observed cleavage patterns, including those not conforming to the even-electron rule, based on literature records and database statistics (Figure 1a). Our strategy for MS/MS annotations allows for exceptions and we admit that they may not cover all fragment ions including the formation of ion-neutral complexes, charge-remote fragmentations,[21] and ongoing isomerizations.

Examples of Peak Annotation Using HR Rules. In this section we present MS/MS annotations of three compounds to demonstrate the usage of HR rules (Figure 2). In all cases we used product ions that exceeded the base peak by 2.5% for the annotations. Moreover, key fragmentations were also checked for the enthalpy decrease between precursor and product ion pairs using the semiempirical MOPAC (Molecular Orbital PACkage) program[18] to ascertain that the fragment ions were stable enough to be monitored in mass spectrometers. The detailed fragmentation schemes with the enthalpy calculation are provided in Supporting Information, Figure S1 and Table S2.

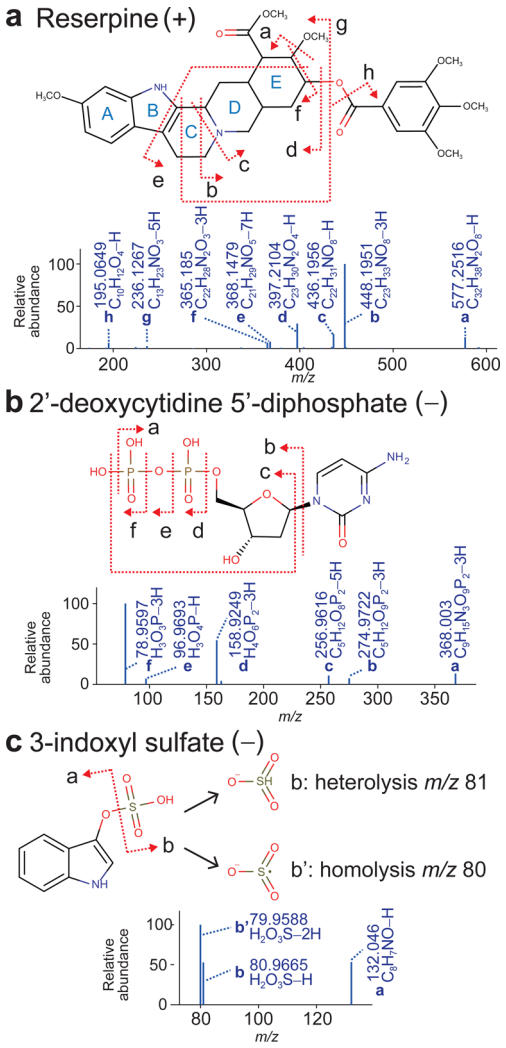


Figure 2. Predicted fragmentation patterns by HR rules. MS/MS spectra of reserpine (a) was measured in the positive and the MS/MS spectra of 2′-deoxycytidine 5′-diphosphate (b) and 3-indoxyl sulfate (c) in the negative ion mode. The arrows indicate bond cleavages. Associated formulas with rearranged hydrogens are shown under each structure for each labeled cleavage.

First, an indole alkaloid reserpine is shown to explain the interpretation of ring cleavages in positive ion mode (Figure 2a). The fragment ions *a*, *d*, and *h* were explained by rule P1. Fragment *f* from two C-bond cleavages was resolved by the

DOI: 10.1021/acs.analchem.6b00770
Anal. Chem. 2016, 88, 7946−7958

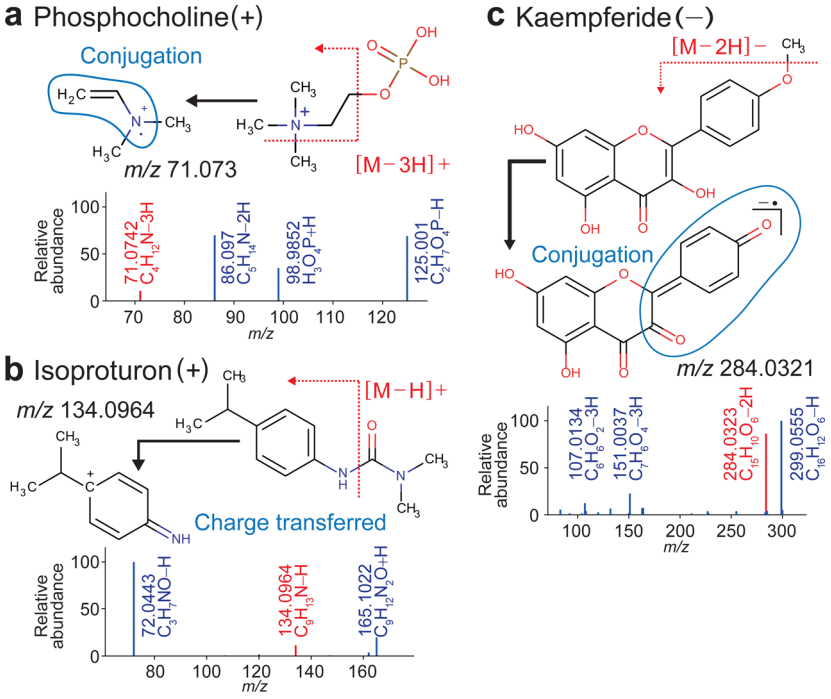


Figure 3. Radical and the charge-transferred ions that are semiresolved by HR rules. MS/MS spectra of phosphocholine (a) and isoproturon (b) were measured in the positive and the MS/MS spectra of kaempferide (c), in the negative ion mode. Depicted in red are the semiresolved peaks in the MS/MS spectra. The range of (hyper-) conjugation confirmed by the MOPAC program is encircled in blue. The remaining peaks in blue were resolved by the rearrangement rules.

combination of rules P1 and P4, as was fragment ion *b*. On the other hand, fragment ion *c* was annotated as the result of C- and N-bond cleavages. The rule combination is P2 and P4 if the first cleavage is the N−C bond, but it is P1 and P3 if the first is the C−C bond. The rules do not tell us the order of fragmentation; only the number of rearranged hydrogens. Fragment ion *e* was the result of two ring cleavages denoted as B and E. The hydrogen rearrangements were the combination of rule P1 applied once, and rule P3 applied three times; the final annotation was $[M - 7H]^+$. Here again, the breaking order was unknown. Likewise, fragment ion *g* was annotated as $[M - 5H]^+$. As in this reserpine example, ring fission often resulted in double-bond formation (rule P4) and contributed to the stabilization of fragment ions by (hyper-) conjugation.

The second example, 2′-deoxycytidine 5′-diphosphate, is shown to explain the behavior of phosphates (Figure 2b). Fragments *m/z* 79 (PO_3^-) and 97 ($[H_3PO_4 - H]^-$) were explained by rule N2. The ion *m/z* 79 was the result of a two-hydrogen loss from phosphate; this is characteristic in CID fragmentations of di- or triphosphate cleavages (fragments *d* and *f* in the spectrum).[22] Although the ion *m/z* 79 may be interpreted as the result of dehydration from *m/z* 97, this interpretation does not hold for the spectrum of guanosine 5′-diphosphoglucose (Supporting Information, Figure S2). Fragment ion *c* ($[C_5H_{12}O_8P_2 - 5H]^-$) can be explained by the combination of rules N2 (loss of two hydrogens) and N5 (loss of one hydrogen). Fragments *b* and *e* were also resolved with rules N2 and N1, respectively.

The last example is the homolysis of the S-containing metabolite 3-indoxyl sulfate (Figure 2c). Fragment ions *a* and *b* were explained by the stability of even-electrons as in rule N1. The fragment ion *m/z* 80 (b′: SO_3^-) was resolved as the result of homolysis in rule N3 (loss of one hydrogen). This odd-electron fragment ion is frequently monitored in sulfones and thiols, and indeed, we found homolysis fragmentations to be more frequent than heterolysis in MassBank spectra (Table 1). The enthalpy decrease in homolytic fragments was confirmed by the MOPAC program; the decrease was almost the same as the decrease resulting from heterolysis (Supporting Information, Table S2).

Treatment of Irregular HR Behaviors as Exception. Not all fragments are resolved by the HR rules. In their review of the even electron rule,[17] Karni and Mandelbaum stated "... although the generalization (even-electron rule) may be a helpful guide for the explanation of mass spectral behavior of many organic compounds, the term "rule" seems to be unjustified.". The same applies for our rules. In Table 1, three major exceptions that deviate from the rules can be identified: (1) +1H or +2H cations from C-bond cleavage (160 and 216 counts each), (2) ±0H cations from N-bond cleavage (129 counts), and (3) −1H anions from O-bond cleavage (187 counts). Since the C-cleaved patterns were well-documented by Karni and Mandelbaum, we focused on the exceptions of N- and O-cleaved fragments. Such odd-electron fragments should be stable enough to be monitored in mass spectrometers, and a plausible explanation for their emergence is the formation of double bonds, that is, (hyper-) conjugation.

Phosphocholine exemplifies a frequently appearing odd-electron cation of an N-cleaved fragment (Figure 3a). The fragment *m/z* 71.0742 was recognized as the odd-electron cation derived from two cleavages of N−C and C−O bonds. The double bond was thought to be formed at the C-cleaved terminal (rule P4), stabilizing the N-radical by the conjugation moiety (blue circle in Figure 3a). Another stabilizing mechanism is the charge transfer as shown for isoproturon (Figure 3b). The fragment of *m/z* 134.0964 can move its charge to the adjacent benzene ring and the nitrogen is stabilized by its hyper-conjugation form. In our manual

DOI: 10.1021/acs.analchem.6b00770
Anal. Chem. 2016, 88, 7946−7958

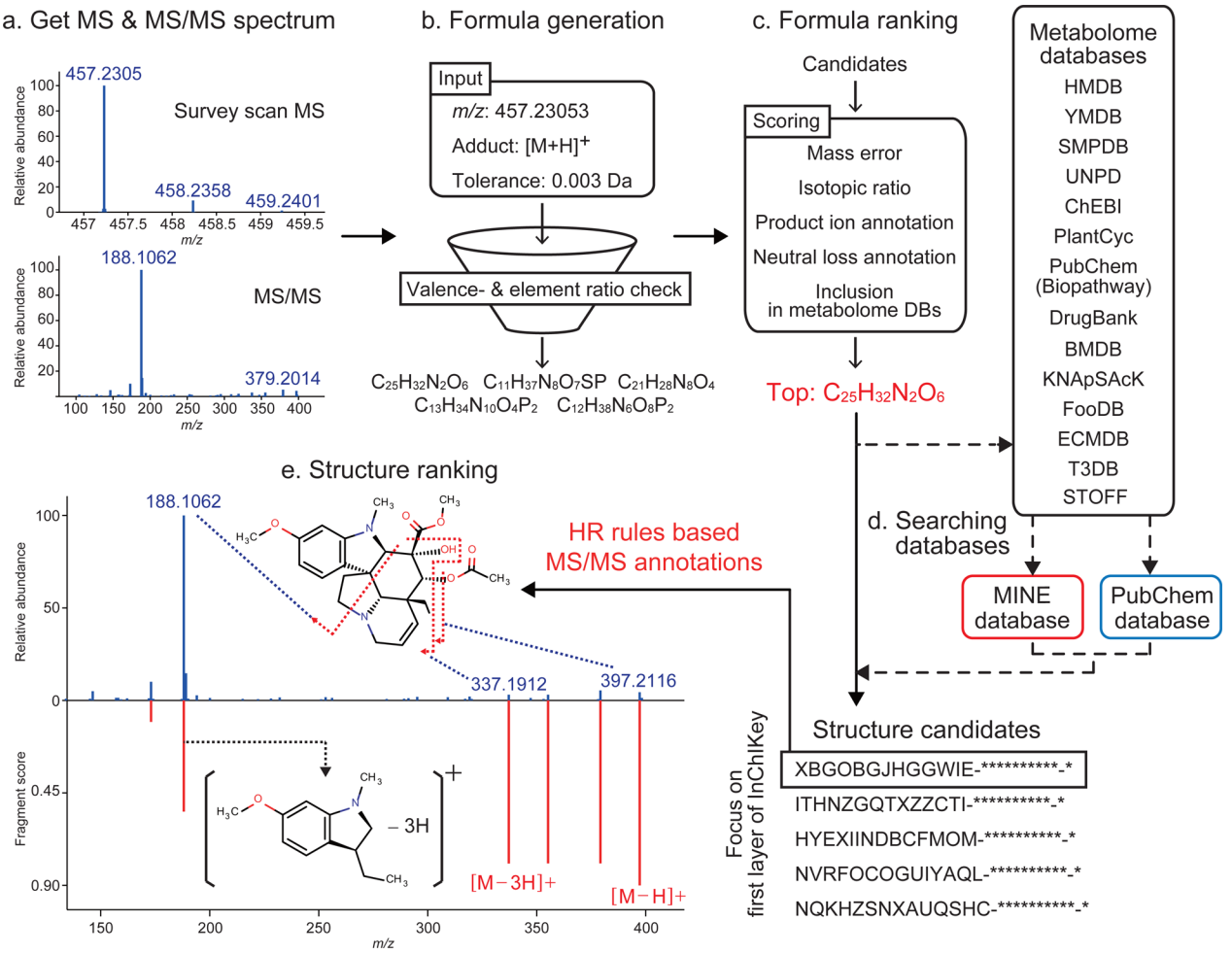


Figure 4. Outline of the MS-FINDER program. (a) Survey MS[1] scans and MS/MS spectra are imported. (b) Formula candidates are generated from the precursor *m/z*, adduct type, and mass tolerance, followed by a filtering with the valence rules and elemental ratios. (c) Candidate formulas are ranked by the sum of five scores. (d) The structure records matching the formula are retrieved from the internal 14 databases, from the MINE database, or from the PubChem repository. (e) Structures are integrated by the first 14 characters of InChIKey and are ranked by the integrated score including the HR rules.

inspection, we could rationalize most exceptions of N-cleaved fragmentation similarly.

The example for an odd-electron anion of an O-cleaved fragment is kaempferide, or 4′-*O*-methylkaempferol (Figure 3c). The fragment was generated by homolysis of the O−C bond whose scheme has been depicted in electron ionization (EI) fragmentation.[17] Its radical anion was stabilized by the broad conjugation moiety (blue circle in Figure 3c). Since flavonoids are typical targets in plant metabolomics, many similar anion radicals from flavonoids are registered in MassBank.

Stable radicals are often monitored but not general enough to be formulated as rearrangement rules because of their dependence on the surrounding atomic environment. Therefore, we treat these irregular behaviors, that is, odd-electron or charge-transferred ions, ad hoc by adding or removing up to two hydrogens. We describe MS/MS peaks as semiresolved when they coincide with the hydrogen-adjusted masses.

Structure Elucidation Using MS-FINDER in Combination with HR Rules. To embody the benefit of HR rules for structure elucidation, we developed the MS-FINDER program, whose workflow is depicted in Figure 4. The program accepts queries of MS and MS/MS spectra in ASCII text or MSP format through a graphical user interface (GUI) (Figure 4a). From the precursor *m/z*, adduct type, and mass tolerance of spectra, candidate formulas are computationally generated and filtered by using valence rules and elemental ratios (Figure 4b). Default settings for adduct types (hydrogen adduct/loss) and mass tolerance can be manually adjusted through the GUI. The candidate formulas are then ranked by mass errors, isotopic ratio, product ions, neutral losses, and presence in a customized "metabolome formula database" (Supporting Information, Table S3) that currently contains 90227 formulas (Figure 4c, see Materials and Methods). For the top-ranking eq 5 by default), matching structures are retrieved from a customized "metabolome structure database" (Supporting Information, Table S4) that currently contains 224663 unique structures, representing data from 14 available metabolome databases (see Supporting Information, SI Manuscript). For elemental formulas that are not found in the metabolome structure database, MS-FINDER searches the candidate either/both in the in silico metabolome expansion database MINE[23] (currently 643307 unique structures) and in the PubChem compound database (Figure 4d). Finally, candidate structures are scored and ranked by the result of in silico MS/MS annotation using the HR rules (Figure 4e). The MS/MS

DOI: 10.1021/acs.analchem.6b00770
Anal. Chem. 2016, 88, 7946−7958

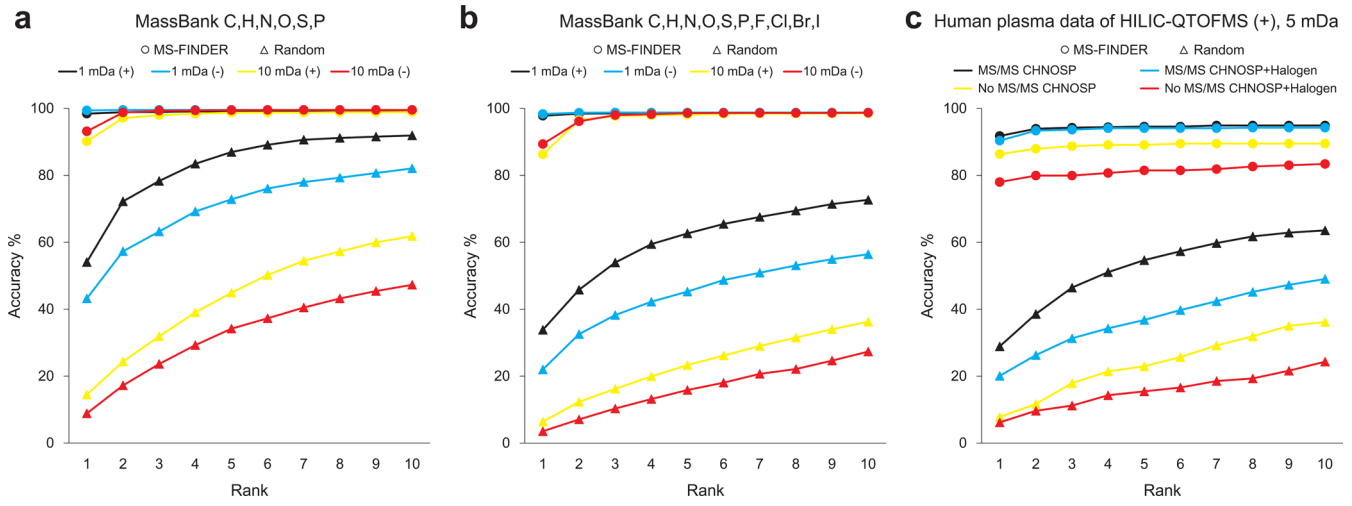


Figure 5. Results of formula prediction. *x* and *y* axes give the ranking and the accumulated percentages of total records. (a) Result of the MassBank validation set when the target elements are set to CHNOPS. (b) Result of the MassBank validation set when the target elements are set to CHNOPS plus halogen atoms. Evaluation was performed with two types of mass tolerances, i.e., 1 and 10 mDa, in positive and negative mode records. (c) Result of the human plasma data set obtained from HILIC-ESI(+)-QTOF MS/MS with 5 mDa mass tolerance. Two types of records were tested: precursor ions with and without MS/MS records. In each panel, "random" indicates a random picking from the candidates as the baseline performance.

scoring is based on mass errors, bond dissociation energy (BDE) differences, a penalty for fragmentation linkage discrepancies, and a penalty for violating HR rules for ranking the candidates (see Materials and Methods for the detail). We note that the workflow is not a de novo prediction, but a constraint-based filtering method for selecting structure candidates. Stereoisomers are not distinguished in this structure search. The program is written in C# language and runs on Windows OS (.NET Framework 4.0 or later; RAM: 8.0 GB or more). It is freely available from the RIKEN PRIMe Web site (http://prime.psc.riken.jp/). We shall explain its performance step by step.

Accuracy of Formula Predictions. Structure annotation starts by obtaining the correct elemental formula for a given spectrum. We assessed the efficiency of our formula prediction from mass spectra based on reference spectra deposited in MassBank (Supporting Information, Table S5). MS-FINDER was first tested on 2708 MassBank MS/MS records in ESI(+) and 1299 records in ESI(−) for molecules that consisted only of CHNOPS elements (Figure 5a). With a mass tolerance of 1 mDa, achievable by modern accurate mass instruments, the correct formula ranked first for 98.7% of the records (51 false negatives). Among the false negative formulas, 30 were outside the search window of permitted elemental ratio. The remaining 21 formulas were listed within the top 7 hits for ESI(+) and top 3 hits for ESI(−) records. With a mass tolerance of 10 mDa, 91.1, 98.4, and 99.1% of the correct formula ranked within the top 1, top 3, and top 10 hits, respectively. This high accuracy became possible because the program prioritizes hits that are also found in our metabolome formula database (Supporting Information, Figure S3). In other words, the formula prediction benefits from using a highly curated formula database.

Next, we used the MassBank records for compounds that bear halogen atoms (F, Cl, Br, and I; Figure 5b). Additional 754 records in ESI(+) and 259 records in ESI(−) were included. With a mass tolerance of 1 mDa, 98.0% and 98.7% of the correct formula ranked within the top 1 and top 10 records, respectively. However, when using the mass tolerance of 10 mDa, the rate clearly worsened, yielding 87.2, 97.8, and 98.6% accuracy for the top 1, top 3, and top 10 hits, respectively. This result strongly suggests the importance of checking isotopic ratios when searching chemical formulas that include chlorides and bromides, especially for spectra obtained from medium-resolution mass spectrometers.

We then applied MS-FINDER for formula predictions using data of a human blood plasma cohort study, acquired by hydrophilic interaction chromatography-ESI(+)-QTOF MS with data dependent MS/MS fragmentations (Figure 5c and Supporting Information, Table S6). In this study, 936 peaks have been manually annotated (677 by MS/MS spectra matching, and 259 by accurate mass and retention time matching with 169 authentic compounds from a total of 6 samples; also see Supporting Information, SI Manuscript). The hand-annotated metabolites served as true-positives for testing the automatic formula determination in MS-FINDER. Within a mass tolerance of 5 mDa, 91.8, 94.3, and 94.9% of the correct formula of CHNOPS metabolites ranked within the top 1, top 3, and top 10 hits, respectively. If MS/MS information was unavailable, the accuracy rates dropped to 86.4, 88.7, and 89.5% for the top 1, top 3, and top 10 hits, respectively. When the target elements were expanded to include halogens, the percentages within the top 1, top 3, and top 10 hits were 90.4, 93.7, and 94.3% for molecules with MS/MS information, and 78.0%, 79.9%, and 83.4% for molecules without MS/MS, respectively. For successful ranking and correct formula calculations, presence in the metabolome formula database proved to be highly important (Supporting Information, Figure S3), as did the use of MS/MS information and rationalizing (or penalizing) fragmentations through HR rules.

Example of Structure Identification. Our hypothesis is that the number of fragment ions resolved by HR rules are larger for the correct structure than for incorrect ones, as indicated from Table 1 (MassBank) and Supporting Information, Table S1 (GNPS). To reflect this assumption, we introduce three levels of annotation (resolved, semiresolved, and unresolved) by applying the HR rules and scoring each fragment ion. In addition, we use traditional methods such as the calculation of bond dissociation energies (BDEs) to rank

DOI: 10.1021/acs.analchem.6b00770
Anal. Chem. 2016, 88, 7946−7958

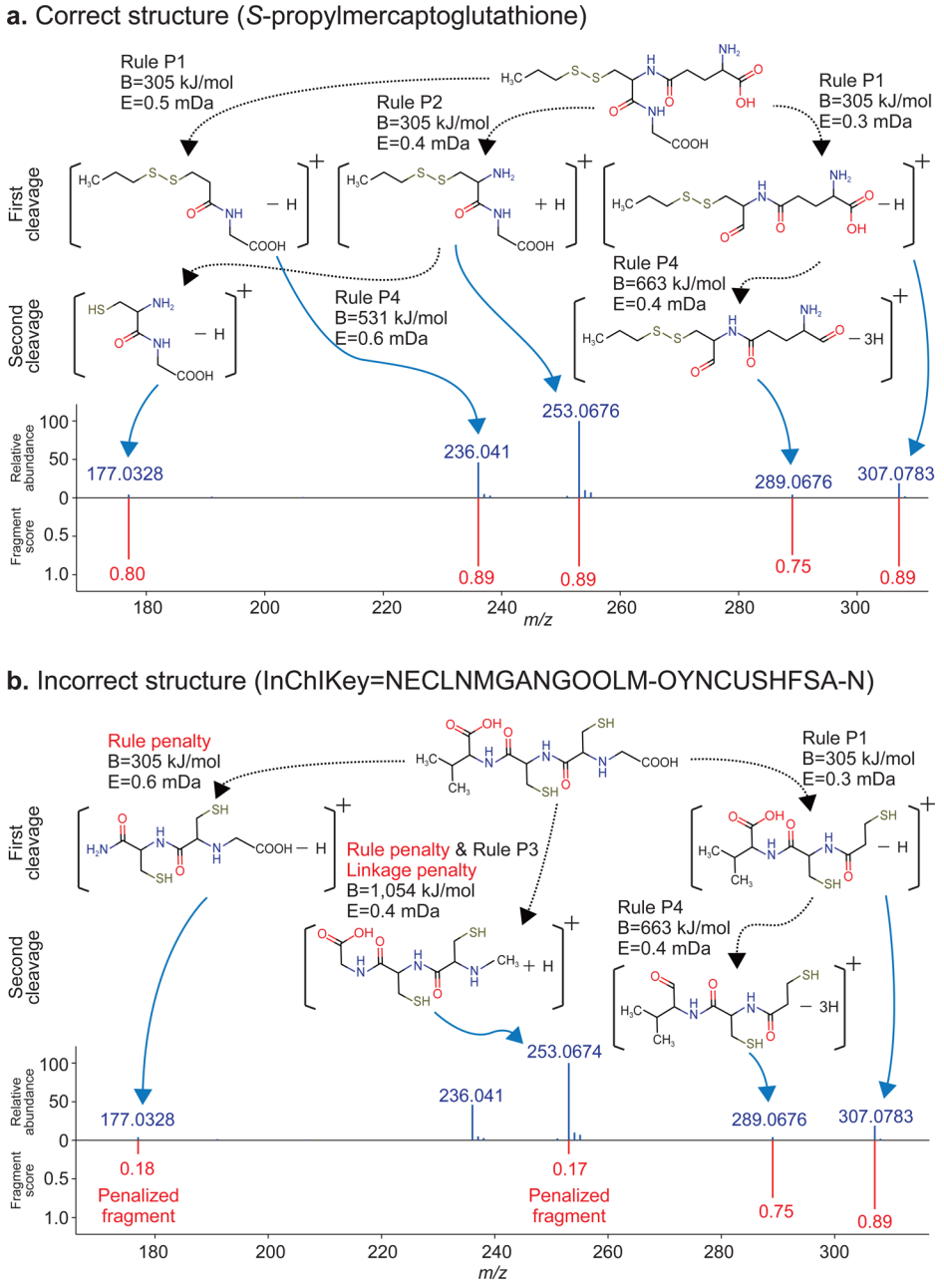


Figure 6. Annotated fragments by the HR rules (red) versus experimental MS/MS spectrum (blue) for (a) *S*-propylmercaptoglutathione and (b) an incorrect structure. Mass error (E) and the accumulated bond dissociation energy (B) are shown with the putative fragmentation scheme. The fragment score is also shown for each predicted peak. In the correct structure (a), all experimental peaks were explained by the rules. In the incorrect structure (b), the fragment for m/z 177.0328 of the C−N bond cleavage was penalized because the exact mass corresponds to no hydrogen transfer instead of two hydrogen transfers by rule P2 in Table 1. Similarly, the fragment for m/z 253.0674 of the C−C bond cleavage was penalized because the exact mass corresponds to two hydrogen transfers instead of no hydrogen implied by rule P1. In addition, the precursor ion of this fragment is not included in the experimental spectrum, causing the additional fragmentation-linkage penalty.

predicted fragments,[10,24] and penalties for fragmentation linkage (precursor-product consistency in HR rules).[25]

As an example to highlight our strategy, we use *S*-propylmercaptoglutathione ($C_{13}H_{23}N_3O_6S_2$), found in onion bulbs.[26] Its molecular formula has 54 isomeric structures in the PubChem compound database. Among them, only the correct structure was successfully predicting all experimental MS/MS ions as "resolved" fragment ions by HR rules (P1, P2, and P4 of Figure 1a; Figure 6, top) with the mass tolerance of less than 1 mDa. Moreover, two product ions (m/z 177.0328 and 289.0676) were correctly assigned to their respective precursor ions (m/z 253.0676 and 307.0783), meaning that no fragmentation linkage penalty was needed. In contrary, none of the 53 incorrect isomeric structures could sufficiently explain the observed spectra as resolved. For example, the structure of *N*-[*N*-[*N*-(carboxymethyl)-cysteinyl]-cysteinyl]-valine (Figure 6, bottom; InChIKey = NECLNMGANGOOLM-OYNCUSH-FSA-N) did not predict the ion of m/z 236.041, and the ion of m/z 177.0328 was explained as irregular HR behavior: the result of N−C bond cleavage should follow rule P2 (causing the rule penalty). In addition, the precursor ion of m/z 253.0676

DOI: 10.1021/acs.analchem.6b00770
Anal. Chem. 2016, 88, 7946−7958

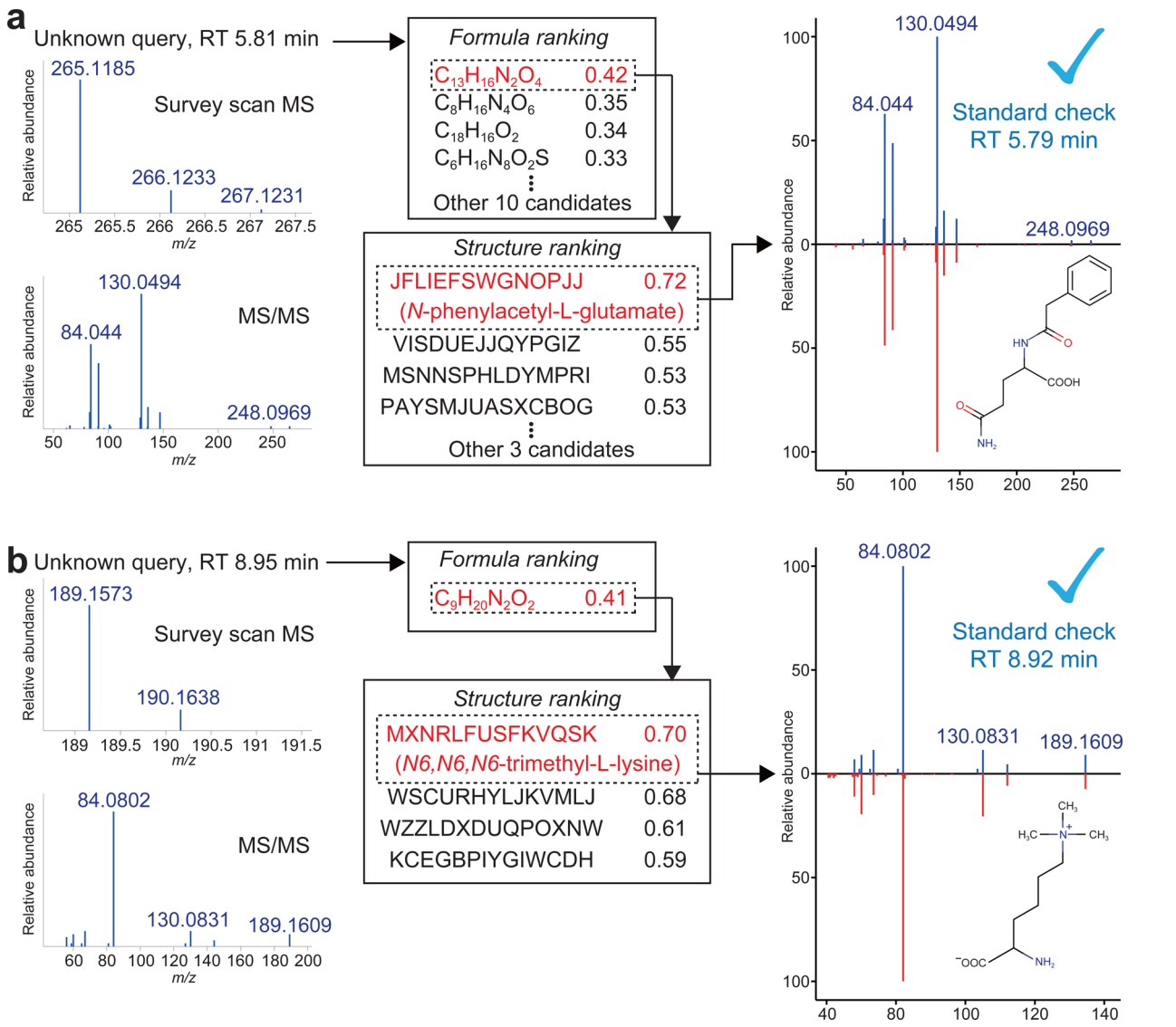


Figure 7. Two examples of compound identification by MS/MS matching in human plasma. MS-FINDER found two new compounds that are absent from MS/MS libraries: (a) *N*-phenylacetyl-L-glutamate and (b) *N6,N6,N6*-trimethyl-L-lysine. Left panels: precursor MS1 and fragment ion MS/MS spectra. Middle panels: formula prediction and structure search results for these spectra, sorted by the MS-FINDER score. Right panels: The experimental MS/MS spectra from human plasma and purchased authentic standards are shown in blue and red, respectively.

was absent in the MS/MS spectrum (causing the linkage penalty).

Moreover, we explored using MS-FINDER for experimental HILIC-ESI(+)-QTOFMS data set of human plasma. The software annotated two compounds whose MS/MS spectra were neither registered in the freely accessible public MoNA spectral repository (MassBank of North America, encompassing spectra from MassBank, HMDB, GNPS, ReSpect, and LipidBlast) nor registered in the licensed spectra libraries Metlin and NIST14. The first correctly annotated compound was *N*-phenylacetyl-L-glutamine, a metabolite that is synthesized in human liver and kidney and excreted in urine (Figure 7a). Its molecular formula $C_{13}H_{16}N_2O_4$ was the top candidate as the result of formula prediction among 14 candidates. The structure ranked as the top among the seven structural isomers for $C_{13}H_{16}N_2O_4$ in the internal structure database. The second example was *N6,N6,N6*-trimethyl-L-lysine, an excretory metabolite from proteolysis (Figure 7b). Within the mass tolerance of 5 mDa, the formula $C_9H_{20}N_2O_2$ was the only candidate and the structure was ranked as the top among four structural isomers. The MS/MS spectra for these compounds (as well as their retention times) coincided with those of subsequently purchased authentic standards. These cases exemplify the accuracy and utility of our software platform.

Overall Accuracy of MS-FINDER. The statistics of structure-selection accuracy was obtained from the same data set used for the formula prediction (Figure 8). For each MS and MS/MS spectrum, five candidate formulas were estimated in this study and all corresponding structures were retrieved from the internal structure database. For formulas that were absent in the internal database, PubChem was used for searching structures: in this study, 50912 of 90332 formulas (56.4%) were derived from PubChem. Testing 5063 MassBank MS/MS spectra within the mass tolerance of 1 mDa for metabolites of CHNOPS elements, the probability of finding the correct structure as the top hit was 64.3%. The percentage that the correct structure was found among the top 3 and top 10 hits was 82.1 and 93.2%, respectively. When increasing the

DOI: 10.1021/acs.analchem.6b00770
Anal. Chem. 2016, 88, 7946−7958

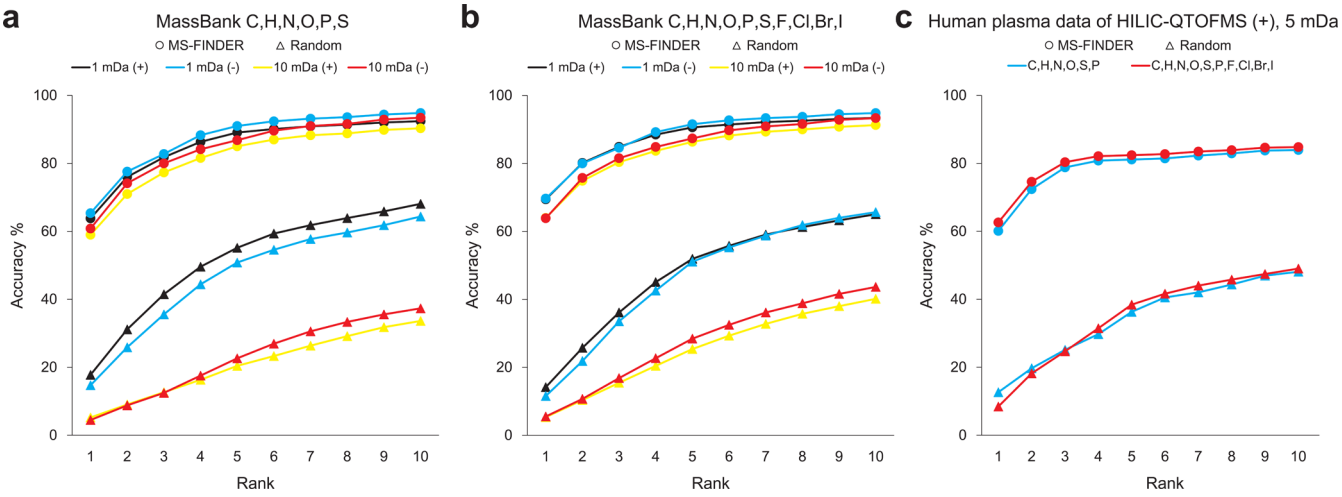


Figure 8. Results of structure selection. *x* and *y* axes show the ranking and the accumulated percentages of total records. (a) Result of the MassBank data when the target elements were set to CHNOPS. (b) Result of the MassBank data when the target elements were set to CHNOPS and halogens. Evaluation was performed with two types of mass tolerance (1 and 10 mDa). (c) Result of the human plasma data obtained from HILIC-ESI(+)-QTOFMS with 5 mDa mass tolerance. Only those records with MS/MS spectra were used. In each panel, "random" indicates a random picking from the candidates as the baseline performance.

Table 2. Comparison of Structure Elucidation Programs[a]

	N6,N6,N6-trimethyl-L-lysine *m/z* 189.1574 MS^1 tolerance 5 mDa, 10 ppm MS^2 tolerance 25 mDa, 20 ppm		*N*-phenylacetyl-L-glutamate *m/z* 265.1186 MS^1 tolerance 5 mDa, 10 ppm MS^2 tolerance 25 mDa, 20 ppm		*S*-propylmercaptoglutathione *m/z* 382.1105 MS^1 tolerance 5 mDa, 10 ppm MS^2 tolerance 10 mDa, 10 ppm	
	rank	computational time	rank	computational time	rank	computational time
MS-FINDER (original search)	1	less than 1 s	1	12 s	1	1 min and 1 s
MS-FINDER (from candidates)	1	3 s	1	1 min and 56 s	1	2 min and 8 s
CSI:FingerID (original search)	1	7 s	1	14 s	not found	11 s
MAGMA (from candidates)	13	50 s	11	3 min and 30 s	1	2 min and 46 s
CFM-ID (from candidates)	25	1 min and 37 s	3	3 h 26 min and 13 s	1	11 h 41 min and 47 s
MetFrag (from candidates)	26	23 s	3	8 min and 56 s	1	17 min and 9 s
MIDAS (from candidates)	25	59 s	11	5 h 27 min and 16 s	2	27 h 51 min and 19 s

[a]The term "original search" indicates the result of using their own resources to compute outputs. The term "from candidates" indicates the result of using the same data set we prepared. Total 86, 1496, and 2368 structures were used for the three structures from left to right, respectively.

mass tolerance to 10 mDa, the ratios were 59.6, 78.2, and 91.3% for the top 1, top 3, and top 10 hits, respectively (Figure 8a). The same trend was observed for metabolites with halogens: 69.5, 84.8, and 93.8% for 1 mDa tolerance and 63.9, 80.7, and 91.9% for 10 mDa tolerance (Figure 8b). For the metabolites in the human plasma data sets, we used only a 5 mDa tolerance window. The top 1, top 3, and top 10 ratios were 60.1, 78.8, and 83.9% for CHNOPS metabolites, and with halogens, the ratios became 62.6, 80.4, and 84.8%, respectively (Figure 8c). Overall, the correct structure ranked within the top-3 with 80% accuracy. This result largely remained even when PubChem entries were always searched, (Supporting Information, Figure S4), because our scoring scheme adds a bonus when the first layer of InChIKey (14 characters) was found in the in-house metabolome databases. This result reconfirms the significance of the internal metabolome database. In Figure 8, all accuracy curves quickly increase from rank 1 to rank 3, because most metabolites have structural isomers: the position of double bonds, positional isomers such as *ortho*, *meta*, *para* orientations, or *cis*−*trans* isomers cannot be distinguished in our software. Therefore, we recommend users to verify at least five candidates for accurate identification.

Comparison of MS-FINDER against Other Programs. We compared the performance of MS-FINDER with five other programs: CFM-ID,[12] CSI:FingerID,[13] MetFrag,[24] MIDAS,[25] and MAGMA[27] on the three highlighted compounds in this paper (*N6,N6,N6*-trimethyl-L-lysine, *N*-phenylacetyl-L-glutamate, and *S*-propylmercaptoglutathione; Table 2). The detail of input parameters was described in Supporting Information, SI Manuscript. Since both MS-FINDER and CSI:FingerID use highly curated internal databases for efficiency, they are denoted as "original search" in Table 2. Although *S*-propylmercaptoglutahione was missed by CSI:FingerID, the overall performance including the computational time was similar. One practical difference is that MS-FINDER can deal with both positive and negative ESI modes, while CSI:FingerID focuses on the positive ion mode. Another difference is the flexibility of MS-FINDER, which can use a user-defined structure database.

To compare the performance on the same information background, we prepared 86, 1496, and 2368 isomeric structures for *N6,N6,N6*-trimethyl-L-lysine, *N*-phenylacetyl-L-glutamate, and *S*-propylmercaptoglutathione, respectively. Using identical mass tolerance thresholds for all programs, the selection result from the candidate structures showed that MS-FINDER ranks among the best in a short computational time. Deeper comparison is, however, needed that could deviate from default parameter settings and that could include

DOI: 10.1021/acs.analchem.6b00770
Anal. Chem. 2016, 88, 7946−7958

statistical comparisons across thousands of spectra. We must also note that the computational cost of web-application tools (CSI:FingerID, MetFrag, and MAGMA) depends on network and server conditions, varying over time. Still, the current comparison results were similar to previous reports[13] and suffices to grasp the tendency of each approach.

The HR rules can be implemented in other programs. It may be argued that the rules are implicitly reflected in machine-learning approaches, but the importance of embodying the fragmentation knowledge as the rules cannot be more emphasized.

CONCLUSION

We proposed nine rules of hydrogen rearrangement for CNOPS elements as the first empirically validated rule set to understand MS/MS fragmentation for small molecules. Odd-electron fragments are semiresolved within ±2 differences in hydrogens. The novelty of our rules lies in the formulation of P- and S-containing fragments whose behaviors are different from fragments containing N and O. The statistics of MassBank MS/MS spectra indicated that our procedure can explain close to 80% of the MS/MS fragment ions in low-energy CID. Our rules thus facilitate the interpretation of cleaved bond positions and help to clarify the mechanism(s) of further isomerization that yields the remaining 20% of product ions.

We developed MS-FINDER program to facilitate formula predictions and isomer structure selections from MS/MS spectra in a seamless way in a single package. Testing human plasma metabolome data sets obtained by HILIC-QTOFMS, the accuracy of formula prediction was 91.8% and over 80% of manually identified structures were correctly selected within the top 3 candidates from the internal structure databases. Moreover, two new compounds were identified without their standard spectra in the reference library. Our final goal is to interpret all MS/MS fragments with theoretical, not empirical, criteria and to identify as many unknown metabolites as possible.

MATERIALS AND METHODS

Construction of Hydrogen Rearrangement Statistics. To obtain the statistics of hydrogen rearrangements, MS/MS records of MassBank were filtered by their peak m/z values. Only those peaks whose m/z values were within 10 mDa of their theoretical mass and whose intensity exceeded 10% of their base peaks were used. The curation detail was described in Supporting Information, SI Manuscript. The assignment of fragment ions was performed in a combinatorial fashion for each molecular structure as follows: (1) up to two bonds were considered broken for a tree-like structure, and up to four bonds were considered when two rings were dissociated (one ring fission was treated as breaking two bonds); (2) hydrogen rearrangement at each bond cleavage was considered within ±2 hydrogens; (3) fragment ions were assigned for observed masses in the following order to specify the most appropriate one from the candidates that match the m/z value within the mass tolerance:

1. Fragments from the first cleavage that follow the even-electron rule.
2. Fragments from the first cleavage that differ up to two hydrogens from a neutralized substructure.
3. Fragments from the second cleavage that follow the even-electron rule.
4. Fragments from the second cleavage whose precursors are also assigned in higher m/z area.
5. Fragments of minimum mass errors.

Strategy for the Assignment of In Silico Fragment Ion by HR Rules. Computational MS/MS annotation was performed by the HR rules and not by the even-electron rule, as follows:

1. Fragments from the first cleavage that follow the HR rules (regular HR behaviors).
2. Fragments from the first cleavage that differ up to two hydrogens (irregular HR behaviors).
3. Fragments from the second cleavage that follow the HR rules.
4. Fragments from the second cleavage whose precursors are also assigned in higher m/z area.
5. Fragments of minimum mass errors.

MOPAC Program for Verifying Stable Fragments. The MOPAC2012 program with PM7 parametrization was manually applied for each ionic structure to check its stability. The program computes the heat of formation from four parameters: the electronic energy, the nuclear−nuclear repulsion energy, the energy to strip off all valence electrons, and the total heat of atomization of all atoms. The latter two parameters were set empirically for each element using the default parameter set. Ionic substructures in the MOL format are available on request from the authors.

MS-FINDER: Computational Generation of Molecular Formulas. Formula candidates were computationally generated from the precursor m/z and the adduct type for a user-defined mass tolerance. The current program accepts a total of 11 elements including C, H, N, O, P, S, F, Cl, Br, I, and Si, although Si is not mentioned in this paper. The valence state and elemental ratio reduce the number of formula candidates.[28] The valence check was based on two equations using the total number of atoms (TA), the total number of atoms having odd valences (OV), and the sum of valences (SV) as follows: (1) OV or SV is even; (2) SV is greater than or equal to 2× (TA − 1). The valence of hetero atoms, that is, N, O, P, and S, were calculated as 3, 2, 5, and 6, respectively. The elemental ratio, such as the hydrogen/carbon balance, was based on our statistics which come from available metabolome databases (Supporting Information, Tables S3 and S7).

MS-FINDER: Ranking of Formula Candidates. The ranking of formulas was based on the total sum of five diagnostic scores on mass error, isotopic ratio, formula assignment to product ions, neutral loss searching, and existence in metabolome databases. Each score was standardized from zero (no similarity) to one (complete match), and the total formula score F ranges from 0 to 5. Each diagnostic score is calculated as follows.

The score of mass error was based on the Gaussian function as in the previous report.[4]

$$\text{mass error score} = \exp\left[-0.5\left(\frac{\text{mass}_{\text{exp}} - \text{mass}_{\text{theor}}}{\delta}\right)^2\right] \quad (1)$$

The background hypothesis of the above equation is that the difference between experimental and theoretical values follows the Gaussian distribution. (The tolerance value, i.e., the standard deviation is user-defined.)

The similarity of isotopic ratio was calculated for 1 (M + 1) and 2 Da (M + 2) shifts from the monoisotopic ion (M). The

DOI: 10.1021/acs.analchem.6b00770
Anal. Chem. 2016, 88, 7946−7958

isotopic abundance was theoretically calculated by the nominal binning. The experimental ions of M + 1 were summed within (precursor m/z + 0.997 − δ, precursor m/z + 1.006 + δ) so that all isotopic ions derived from C, H, N, O, P, S, F, Cl, Br, I, and Si were covered. The ions of M + 2 were also summed within (precursor m/z + 1.994 − δ, precursor m/z + 2.013 + δ). (The tolerance value is user-defined.) After the isotopic abundances were normalized by the monoisotopic ion, the similarity was calculated as follows.

$$\text{istopic ratio score} = \frac{\exp\left[-0.5\left(\frac{M+1_{exp}-M+1_{theor}}{\delta}\right)^2\right] + \exp\left[-0.5\left(\frac{M+2_{exp}-M+2_{theor}}{\delta}\right)^2\right]}{2} \quad (2)$$

The terms of M + 1_{exp} and M + 2_{exp} show the experimental ion abundances of M + 1 and M + 2, respectively. The terms of M + 1_{theor} and M + 2_{theor} show the theoretical abundances of M + 1 and M + 2, respectively. (The standard deviation is user-defined.)

The score for formula assignment was the rate of product ion assignment. After assigning elemental formulas for all product ions, the isotopic ions were excluded since the reliability of their isotopic ratio depended on the precursor isolation settings. The deisotoping of product ions was conducted with the following criteria: (1) isotope peak must be smaller than 1.5-fold of the theoretical isotopic abundance given that the elements of the precursor ion were carbon only, and (2) the m/z of isotopes is found within the user-defined mass tolerance.

For neutral losses, mass differences between all pairs of product ions were checked, and only the formulas in our neutral loss database (total 135 neutral losses from MS2Analyzer[29]) were used as neutral losses. The assigned percentage of neutral losses for all differences was utilized as the evaluation score.

The score of database existence was discrete. If the formula was included in any of the metabolome databases that we curated (Supporting Information, Table S3), 0.5 was used as the evaluation score, otherwise 0. To this value, the number of databases that include the formula, standardized by 0.5, was added.

MS-FINDER: Searching of Structure Candidates. The structural isomers for a given formula were retrieved from the internal database curated from total 14 metabolome databases (Supporting Information, Table S4). If no entry was found, the program provides two options: searching the metabolic in silico network expansion database MINE[23] or/and using the PubChem PUG REST service to obtain structure data (the download limit can be specified). Structures were merged according to the first 14 characters of their InChIKeys to remove the stereo information. The structure with the largest number of synonyms, obtained by PubChem Identifier Exchange Service, was chosen as a representative structure.

MS-FINDER: Ranking of Structure Candidates. Structures were ranked by the scores on the basis of mass accuracy, bond dissociation energies (BDE), penalty of fragmentation linkage, and penalty of HR rules. First, the in silico MS/MS annotation is executed by the method of Strategy for the Assignment of In Silico Fragment Ion by HR Rules section. Unresolved mass peaks by the HR rules were penalized by the score, P_h (0.2 by default). Mass error penalty, M, was calculated by the Gaussian function as described in the section of formula ranking. The penalty of BDE was calculated by the following equation,

$$B = \sqrt{1 - \frac{B_f}{B_m}} \quad (3)$$

where B_m is the maximum BDE in all possible in silico fragmentation of the structure and B_f is the BDE value necessary to produce the fragment. BDE values for different bonding patterns were taken from the internal BDE dictionary. The penalty score of fragmentation linkage P_t (0.5 by default) was introduced to prioritize precursor−product pairs that were both monitored in experimental MS/MS spectra.[25]

The score for each fragment ion was calculated as follows.

$$S_f = M \times B \times P_t \times P_h \quad (4)$$

The penalties of P_t and P_h in this study were set to 0.5 and 0.2, respectively. Then, the sum of all scores for assigned fragments are divided by the number of MS/MS peaks: $(1/n)\sum_{f=1}^{n} S_f$, where n stands for the number of MS/MS peaks. Cumulative fragmentation scores were normalized to 4, and a metabolome database presence score of up to 1 was added to form the overall structure weighting score S, ranging from 0 to 5. The database score was 0 if absent from the in-house metabolome databases and 0.5 + 0.5 × (the number of databases that contains the first layer of InChIKey)/(total number of databases) otherwise. The final ranking of molecular structures was determined by the sum of the formula score F (as explained in the previous section) and the structure score S.

$$T = a \times F + b \times S \quad (5)$$

The coefficients a and b were set to 1 in this study.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.6b00770.

SI Manuscript: (1) The curation detail of MassBank, GNPS, and human plasma data sets, (2) the curation detail of metabolome databases, and (3) the detail of program comparisons (PDF).

Figure S1: Putative fragmentation schemes described in Figure 2 (PDF).

Figure S2: Peak annotation result of guanosine 5'-diphosphoglucose MS/MS spectra and the putative fragmentation scheme (PDF).

Figure S3: Result of formula prediction with the result of database score only. x and y axes give the ranking and the accumulated percentages of total records. (a) Result of the MassBank validation set when the target elements are set to CHNOPS. (b) Result of the MassBank validation set when the target elements are set to CHNOPS plus halogen atoms. (c) Result of the human plasma data set obtained from HILIC-ESI(+)-QTOF MS/MS with 5 mDa mass tolerance. In each panel, "random" indicates a random picking from the candidates as the baseline performance. In addition, "database count" indicates the result derived from the database score only, which is also used as the baseline performance to explain the effect of meta-data scores (PDF).

Figure S4: Result of structure selection for PubChem repository. Structure data for all formula candidates were

DOI: 10.1021/acs.analchem.6b00770
Anal. Chem. 2016, 88, 7946−7958

retrieved from PubChem repository. The data set was the same as that of Figure 8. (a) Result of the MassBank data when the target elements were set to CHNOPS. (b) Result of the MassBank data when the target elements were set to CHNOPS and halogens. (c) Result of the human plasma data obtained from HILIC-ESI(+)-QTOFMS with 5 mDa mass tolerance (PDF).
Table S1: Statistics of hydrogen rearrangements for CNOPS elements in GNPS (PDF).
Table S2: Peak annotation details of Figure 2 (PDF).
Table S3: Internal formula database from 14 metabolome databases (XLSX).
Table S4: Internal structure database from 14 metabolome databases (XLSX).
Table S5: 5063 MassBank records used in this study (XLSX).
Table S6: 936 human plasma records used in this study (XLSX).
Table S7: Formula element statistics of formula databases to determine the practical elemental ratios (XLSX).
Table S8: In-house HILIC-ESI(+)-QTOFMS library of human blood plasma metabolites (XLSX).

AUTHOR INFORMATION

Corresponding Authors

*E-mail: arita@nig.ac.jp.
*Tel.: +1-530-754-8258. E-mail: ofiehn@ucdavis.edu.

Author Contributions

H.T., R.N., K.S., O.F., and M.A. designed the research. H.T. and M.A. created the hydrogen rearrangement rules. H.T. implemented the MS-FINDER program. H.T. and D.Y. discussed the program architecture for speed and graphical user interface. H.T., R.N., K.S. and M.A. designed the fragmentation scheme. H.T. and T.K designed the formula prediction and in silico fragmentation. T.C. performed the experiment for human plasma, its compound identification, and novel compound discoveries. W.T. analyzed the putative fragmentation scheme and calculated its enthalpy changes. R.N. and K.S. prepared the spectral data. H.T., O.F., and M.A. thoroughly discussed this project and wrote the manuscript. The other authors also contributed to the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the NSF-JST Strategic International Collaborative Research Program (SICORP) for JP-US Metabolomics; the Database Integration Coordination Program of the National Bioscience Database Center; Grant-in-Aid for Scientific Research on Innovative Areas 15H05897; the JST-Core Research for Evolutionary Science and Technology (JST-CREST); and partially the Japan Advanced Plant Science Network. H.T. was supported by a grant-in-aid for scientific research (C) 15K01812. O.F. and T.K. were supported by the corresponding Grant NSF MCB 1139644. T.K. was also supported by NSF CBET 1438211, NIH 2R01HL091357-05, and AHA 15SDG25760020. The blood plasma cohort data were acquired as a part of the P20 Program NIH HL113452 Grant supported by National Institutes of Health (NIH). Mass spectral recording and annotations were supported by Grant NIH DK097154 to O.F. We appreciate the support of the QTOFMS instrument through NIH Grant S10-RR031630. A special thanks to Prof. Shigehiko Kanaya (NAIST) and Plant Metabolic Network (PMN) for giving us the KNApSAcK MOL files and PlantCyc database, respectively. Finally, we appreciate Dr. Hisao Tanaka for active discussions about hydrogen rearrangements.

REFERENCES

(1) Fiehn, O. *Plant Mol. Biol.* **2002**, *48*, 155−171.
(2) Patti, G. J.; Yanes, O.; Siuzdak, G. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 263−269.
(3) Benton, H. P.; Ivanisevic, J.; Mahieu, N. G.; Kurczy, M. E.; Johnson, C. H.; Franco, L.; Rinehart, D.; Valentine, E.; Gowda, H.; Ubhi, B. K.; et al. *Anal. Chem.* **2015**, *87*, 884−891.
(4) Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M. *Nat. Methods* **2015**, *12*, 523−526.
(5) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; et al. *J. Mass Spectrom.* **2010**, *45*, 703−714.
(6) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; et al. *Nucleic Acids Res.* **2007**, *35*, 521−526.
(7) Afendi, F. M.; Okada, T.; Yamazaki, M.; Hirai-Morita, a.; Nakamura, Y.; Nakamura, K.; Ikeda, S.; Takahashi, H.; Altaf-Ul-Amin, M.; Darusman, L. K.; et al. *Plant Cell Physiol.* **2012**, *53*, 1−12.
(8) Kind, T.; Liu, K.; Lee, D. Y.; Defelice, B.; Meissen, J. K.; Fiehn, O. *Nat. Methods* **2013**, *10*, 755−758.
(9) Wenk, M. *Nat. Rev. Drug Discovery* **2005**, *4*, 594−610.
(10) Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S. *J. Cheminf.* **2016**, *8*, 3.
(11) Heinonen, M.; Rantanen, A.; Mielikäinen, T.; Kokkonen, J.; Kiuru, J.; Ketola, R. A.; Rousu, J. *Rapid Commun. Mass Spectrom.* **2008**, *22*, 3043−3052.
(12) Allen, F.; Greiner, R.; Wishart, D. *Metabolomics* **2015**, *11*, 98−110.
(13) Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 12580−12585.
(14) Bowen, R. D. *Acc. Chem. Res.* **1991**, *24*, 364−371.
(15) Paizs, B.; Suhai, S. *Mass Spectrom. Rev.* **2005**, *24*, 508−548.
(16) Nakata, H. *Eur. Mass Spectrom.* **1999**, *5*, 411−418.
(17) Karni, M.; Mandelbaum, A. *Org. Mass Spectrom.* **1980**, *15*, 53−64.
(18) Stewart, J. J. P. *MOPAC2012*; Stewart Computational Chemistry; Colorado Springs, CO, U.S.A., 2012 (downloaded on June 3, 2015); http://openMOPAC.net.
(19) Sun, M.; Dai, W.; Liu, D. Q. *J. Mass Spectrom.* **2008**, *43*, 383−393.
(20) Rochfort, S. J.; Trenerry, V. C.; Imsic, M.; Panozzo, J.; Jones, R. *Phytochemistry* **2008**, *69*, 1671−1679.
(21) Cheng, C.; Gross, M. L. *Mass Spectrom. Rev.* **2000**, *19*, 398−420.
(22) Edelson-Averbukh, M.; Pipkorn, R.; Lehmann, W. D. *Anal. Chem.* **2006**, *78*, 1249−1256.
(23) Jeffryes, J. G.; Colastani, R. L.; Elbadawi-Sidhu, M.; Kind, T.; Niehaus, T. D.; Broadbelt, L. J.; Hanson, A. D.; Fiehn, O.; Tyo, K. E. J.; Henry, C. S. *J. Cheminf.* **2015**, *7*, 44.
(24) Wolf, S.; Schmidt, S.; Müller-Hannemann, M.; Neumann, S. *BMC Bioinf.* **2010**, *11*, 148.
(25) Wang, Y.; Kora, G.; Bowen, B. P.; Pan, C. *Anal. Chem.* **2014**, *86*, 9496−9503.
(26) Nakabayashi, R.; Sawada, Y.; Yamada, Y.; Suzuki, M.; Hirai, M. Y.; Sakurai, T.; Saito, K. *Anal. Chem.* **2013**, *85*, 1310−1315.
(27) Ridder, L.; van der Hooft, J. J. J.; Verhoeven, S. *Mass Spectrom.* **2014**, *3*, 1−7.
(28) Kind, T.; Fiehn, O. *BMC Bioinf.* **2007**, *8*, 105.
(29) Ma, Y.; Kind, T.; Yang, D.; Leon, C.; Fiehn, O. *Anal. Chem.* **2014**, *86*, 10724−10731.

DOI: 10.1021/acs.analchem.6b00770
Anal. Chem. 2016, 88, 7946−7958