Journal of proteome research

Article

pubs.acs.org/jpr

Accurate Identification of Unknown and Known Metabolic Mixture Components by Combining 3D NMR with Fourier Transform Ion Cyclotron Resonance Tandem Mass Spectrometry

Cheng Wang,[†,∥] Lidong He,[⊥,∥] Da-Wei Li,[‡,∥] Lei Bruschweiler-Li,[‡] Alan G. Marshall,*[,⊥,¶] and Rafael Brüschweiler*[,†,‡,§]

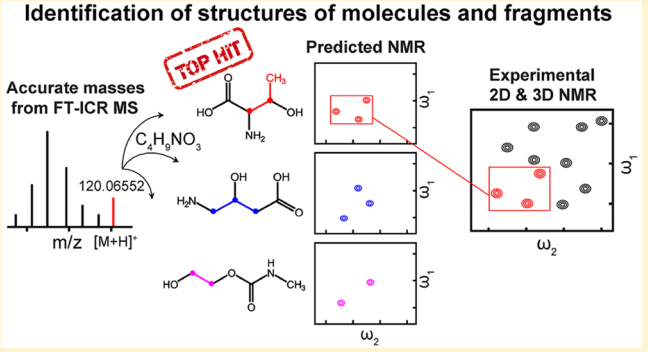
[†]Department of Chemistry and Biochemistry, [‡]Campus Chemical Instrument Center, and [§]Department of Biological Chemistry and Pharmacology, The Ohio State University, Columbus, Ohio 43210, United States

[⊥]Department of Chemistry and Biochemistry, Florida State University, Tallahassee, Florida 32306, United States

[¶]Ion Cyclotron Resonance Program, The National High Magnetic Field Laboratory, Florida State University, Tallahassee, Florida 32310, United States

Supporting Information

ABSTRACT: Metabolite identification in metabolomics samples is a key step that critically impacts downstream analysis. We recently introduced the SUMMIT NMR/mass spectrometry (MS) hybrid approach for the identification of the molecular structure of unknown metabolites based on the combination of NMR, MS, and combinatorial cheminformatics. Here, we demonstrate the feasibility of the approach for an untargeted analysis of both a model mixture and *E. coli* cell lysate based on 2D/3D NMR experiments in combination with Fourier transform ion cyclotron resonance MS and MS/MS data. For 19 of the 25 model metabolites, SUMMIT yielded complete structures that matched those in the mixture independent of database information. Of those, seven top-ranked structures matched those in the mixture, and four of those were further validated by positive ion MS/MS. For five metabolites, not part of the 19 metabolites, correct molecular structural motifs could be identified. For *E. coli*, SUMMIT MS/NMR identified 20 previously known metabolites with three or more ^{1}H spins independent of database information. Moreover, for 15 unknown metabolites, molecular structural fragments were determined consistent with their spin systems and chemical shifts. By providing structural information for entire metabolites or molecular fragments, SUMMIT MS/NMR greatly assists the targeted or untargeted analysis of complex mixtures of unknown compounds.



KEYWORDS: *metabolomics, unknown metabolite identification, NMR-MS hybrid approach, 3D NMR HSQC-TOCSY, COLMAR database*

INTRODUCTION

The large number of different metabolites found in living organisms offers important clues about the chemical underpinning of life, which is the subject of the field of metabolomics. It has been estimated that the human body alone contains over 100 000 different metabolites.[1] Despite ongoing progress in the development of larger metabolomics databases, the identification of unknown metabolites remains a major bottleneck. Traditional approaches for natural product analysis, which are based on the complete physical separation of the compound of interest, are very time-consuming and hence impractical for routine and high-throughput applications. Alternatively, the two primary analytical techniques in metabolomics, namely mass spectrometry (MS)[2−4] and nuclear magnetic resonance (NMR, see below), have been applied separately.

Recently, new approaches have been proposed for the analysis of complex mixture based on combining MS and NMR. Finding ways to synergistically apply the two methods to the same problem has been a challenge due to the high complementarity of their information content. One strategy focuses on subsets of spectroscopic signals that are highly correlated or interdependent with respect to each other across a large number of samples and hence may stem from the same molecule.[5−8] Such correlation analysis can be carried out either for NMR data or direct infusion MS data alone or between the two methods.[9] Groups of signals that have been identified in this way can then be used to deduce information about the molecular structure. These statistical methods are applicable

Received: June 28, 2017
Published: August 10, 2017

DOI: 10.1021/acs.jproteome.7b00457
J. Proteome Res. 2017, 16, 3774−3786

under two conditions, namely that a potentially large pool of samples is available so that statistically meaningful results can be obtained and that the compound of interest shows relatively large modulations of its concentration relative to other metabolites, so that the correlations between signals of the compound are sufficiently large. For applications with smaller sample pools, which can consist of as few as a single sample, alternative approaches have been proposed. For uniformly ^{13}C-labeled mixtures, 2D ^{13}C−^{13}C TOCSY or INADEQUATE experiments permit the tracing of the backbone topology of individual metabolites, thereby providing useful information toward the elucidation of their structure.[10,11] 3D-(H)CCH-TOCSY and COSY spectra of a 60% ^{13}C-labeled rhododendron shrub were used together with quantum-chemical calculations to identify catalogued as well as several uncatalogued metabolites.[12]

We recently introduced approaches that synergistically use NMR and MS for a single sample of a complex mixture at ^{13}C natural abundance for the validation of known compounds and the determination of unknowns. The first approach is the NMR/MS Translator, which translates the metabolites identified from 1D or 2D NMR by database query to accurate masses that are then directly compared with MS of the same sample, thereby providing a methodical validation of metabolites by both NMR and MS.[13,14] The second approach, termed SUMMIT MS/NMR (for "structure of unknown metabolomic mixture components by MS/NMR"),[15] is more complex and more ambitious than some of the other approaches listed because it aims at the determination of the structure of unknown metabolites without the use of NMR or MS databases. On the basis of accurate masses from MS, it generates a pool of possible molecular structures for which NMR chemical shifts are computed and compared directly with the 2D experimental chemical shift data of the mixture spectrum. As a proof-of-principle, it was demonstrated how SUMMIT could determine the correct structures for a finite, well-defined subset of metabolites previously known to exist in *E. coli* cell lysate.[15]

Here, we generalize SUMMIT for the untargeted identification of both known and unknown metabolites by combining ultrahigh-resolution Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS)[16] to assign unique elemental compositions ($C_cH_hN_nO_oP_p$...) with 3D NMR complemented by 2D NMR experiments as the primary source of information for spin-system identification and validation by tandem MS (MS/MS). The NMR information is first queried against the COLMAR NMR database,[17−19] by use of the COLMARm[20] database to identify a maximal number of known metabolites, and thereby assign as many cross-peaks as possible. This step is then followed by SUMMIT, combining MS with NMR data, for the identification of unknowns from the remaining cross-peaks. The approach is first demonstrated here for a model mixture consisting of 25 metabolites. Compared to the original SUMMIT experiments, for which mass measurement was based on time-of-flight mass analysis with an average root−mean−square (rms) mass error of ∼5 ppm (thus limited to metabolites of up to ∼300 Da in mass), the present 9.4 T FT-ICR mass measurements achieve 25-times higher mass accuracy (∼200 ppb rms mass error) and thus allow a more reliable determination of elemental composition for metabolites up to at least 1 kDa in mass.[21] By combining the results with MS/MS analysis, this approach can provide additional disambiguation of the top hits produced by SUMMIT.

METHODS

Identification of Unique Elemental Composition from Ultrahigh-Resolution FT-ICR Mass Spectra

The SUMMIT approach begins from identification of metabolite elemental composition. It has been shown that mass measurement accuracy of ∼200 ppb enables identification of unique elemental composition for organic molecules up to ∼1 kDa in mass.[22] Such mass measurement accuracy for complex mixtures is routinely achieved by 9.4 T FT-ICR MS, for example, resolution and elemental composition assignment for more than 125 000 peaks in a single mass spectrum of a volcanic asphalt sample.[23] Although limited structural information may be derived from elemental composition alone (e.g., Kendrick mass,[24] van Krevelen diagrams,[25] double bond equivalents versus carbon number for individual heteroatom classes,[26] etc.), the most definitive structural information relies on NMR (see below).

Identification of Individual Spin Systems of a Mixture from a 3D ^{13}C−^{1}H HSQC-TOCSY NMR Spectrum

The NMR portion of SUMMIT focuses on the identification of spin systems of individual compounds based on multidimensional ^{13}C−^{1}H and ^{1}H−^{1}H cross-peaks. In particular, the 2D ^{13}C−^{1}H HSQC experiment provides ^{13}C−^{1}H chemical shift correlations between directly bonded ^{1}H and ^{13}C nuclei, and 2D ^{13}C−^{1}H HSQC-TOCSY provides ^{13}C−^{1}H and ^{1}H−^{1}H bond connectivity information. If peak overlaps are absent or rare, the combination of 2D HSQC and 2D HSQC-TOCSY enables the unambiguous extraction of the spin systems of the various mixture compounds. However, in practice the presence of peak overlap will interfere with the accuracy and reliability of spin system extraction. Because the 3D ^{13}C−^{1}H HSQC-TOCSY spectrum is much less prone to peak overlap than its 2D variant, we extract the spin systems directly from the 3D ^{13}C−^{1}H HSQC-TOCSY NMR spectrum. The 3D ^{13}C−^{1}H HSQC-TOCSY experiment provides ^{13}C(ω_1), ^{1}H(ω_2), and ^{1}H(ω_3) correlations and resolves overlap of cross-peaks in the 2D ^{13}C(ω_1)−^{1}H(ω_2) plane by spreading the resonances along the orthogonal ^{1}H(ω_3) dimension, which is the direct ^{1}H detection dimension.[27,28] Comparison of the ^{1}H−^{1}H correlations along ω_3 for each pair of ^{13}C−^{1}H cross-peaks enables one to determine whether or not two ^{13}C−^{1}H cross-peaks belong to the same molecule, thereby drastically reducing the possibility of false spin system identification. In practice, the main concern is the relatively low resolution along the two indirect dimensions that provide the ^{13}C and ^{1}H correlation information to keep the measurement time reasonably short. This problem can be addressed in part by measuring an additional high-resolution 2D ^{13}C−^{1}H HSQC spectrum to complement the ^{13}C and ^{1}H correlation information from the 3D experiment, as done here, or the use of nonuniform sampling methods.[28,29]

The ^{13}C(ω_1)−^{1}H(ω_2) plane of the 3D HSQC-TOCSY spectrum depicts single bond ^{13}C−^{1}H correlations of all molecules in the mixture. To distinguish ^{13}C(ω_1)−^{1}H(ω_2) cross-peaks from different molecules and be able to cluster these cross-peaks into individual spin systems, the analysis of ^{1}H−^{1}H TOCSY transfers along the ω_3 dimension permits one to correlate pairs of ^{13}C(ω_1)−^{1}H(ω_2) cross-peaks and assign them to the same spin system. A prerequisite is that the cross-peaks share the same cross-peaks along ω_3. Specifically, such a pair of 2D cross-peaks, (ω_1',ω_2') and (ω_1'',ω_2''), must then

DOI: 10.1021/acs.jproteome.7b00457
J. Proteome Res. 2017, 16, 3774−3786

share 3D cross-peaks at positions $(\omega_1,\omega_2,\omega_3) = (\omega_1',\omega_2',\omega_2')$, $(\omega_1',\omega_2',\omega_2'')$, $(\omega_1'',\omega_2'',\omega_2'')$, $(\omega_1'',\omega_2'',\omega_2')$. The goal is to find all pairs of 2D cross-peaks that are connected in this manner. These cross-peaks can be considered as edges of a mathematical graph in which the nodes correspond to directly bonded $^{13}C-^{1}H$ spin pairs. Such a graph can then be analyzed in terms of a "maximal clique" analysis, which we recently developed to automatically extract all possible spin systems from TOCSY-type spectra automatically.[30]

Figure 1 depicts a schematic diagram for the generation of spin systems from the 3D $^{13}C-^{1}H$ HSQC-TOCSY NMR

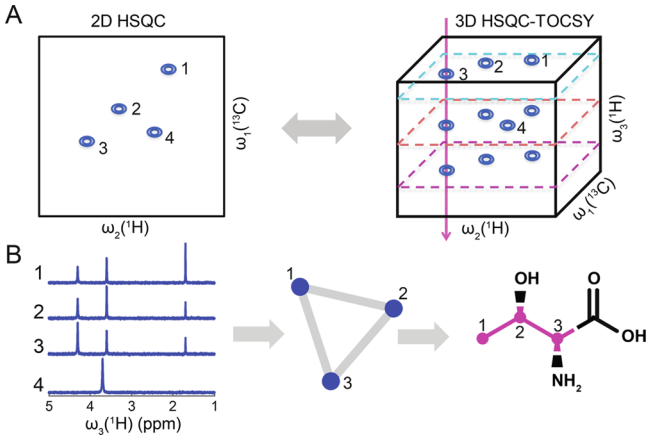


Figure 1. Extraction of spin systems of individual mixture compounds from 3D $^{13}C-^{1}H$ HSQC-TOCSY. Panel A shows the relationship between cross-peaks from the 2D $^{13}C-^{1}H$ HSQC spectrum (left) and the 3D $^{13}C-^{1}H$ HSQC-TOCSY spectrum (right). Panel B illustrates how 1D cross sections along ω_3 (^{1}H) of the 3D HSQC-TOCSY spectrum of (a) yield spin system information, which is extracted by use of a maximal clique approach. Traces 1, 2, 3 show high similarity because they belong to the same spin system consisting of three protons, whereas trace 4 belongs to a separate spin system with a single proton.

spectrum. It should be noted that for spin systems with only one $^{13}C-^{1}H$ pair, the method does not work because the spin system is fully characterized by a single cross-peak in the 3D spectrum. Therefore, peak assignments of one-spin systems need to be performed manually. Similarly, because two-spin systems contain no redundant information, they are more prone to false positives and were considered only for the model mixture but not for *E. coli* cell extract.

After spin systems of individual compounds are extracted, they are refined to minimize the occurrence of false positives. Spin system refinement includes three consecutive steps. First, extracted spin systems are validated by visually checking 2D $^{13}C-^{1}H$ HSQC, 2D $^{1}H-^{1}H$ TOCSY, and 2D $^{13}C-^{1}H$ HSQC-TOCSY NMR spectra. If the extracted spin system was incomplete, expected peaks that are unambiguously observed by visual inspection, but that were missed by the automated procedure, are manually added until the spin system is complete. Second, ^{1}H peak doublets and nearly degenerate ^{1}H resonances are combined into a single chemical shift. For example, CH_2 groups can have two separate proton chemical shifts belonging to the same ^{13}C, but sometimes it is difficult to determine whether two separate peaks stem from a single CH_2 group or from two separate CH groups. Therefore, those spin systems that contain two cross-peaks with the same ^{13}C frequency in the HSQC are merged into a single cross-peak (with a chemical shift taken as the mean of the two proton resonances) for the generation of an alternative spin system candidate, which is added to the list of spin systems. Third, potential extra spins are manually identified and added after comparison of 1D ^{1}H traces along ω_3. For example, for an automatically generated three-spin system, if an additional resonance shows unambiguous connectivities to all three spins, but has not yet been included in the clique, then this spin is manually added, resulting in a new four-spin system. An example for the refinement of spin systems is provided in Figure S1.

Structure Manifold Generation and 2D HSQC NMR Spectra Prediction

Each accurate mass derived from an experimental FT-ICR mass spectrum was compared to the METLIN database to identify the closest matching molecular formula (note that METLIN was used only to search for molecular formulas that are consistent with the FT-ICR-based mass information, but not for molecular structures). Because each molecular formula could correspond to any of several isomers, we then searched the ChemSpider database[31,32] for all structures corresponding to a given molecular formula.

For all molecular structures, 2D $^{13}C-^{1}H$ HSQC spectra are predicted by use of the empirical chemical shift predictor by Modgraph implemented in MestReNova 10.0.1 (Mestrelab Research). HSQC prediction for each molecule takes about 3−10 s with a desktop computer. The ^{13}C chemical shift prediction utilizes a HOSE code algorithm, whereas the ^{1}H chemical shift prediction is based on functional groups which were individually parametrized.[33] Because NMR chemical shift prediction plays a critical role in SUMMIT for identifying the correct compound from a large compound pool, we examined the prediction accuracy for amino acids, organic acids, and carbohydrates contained in a 25-compound model mixture. We compared the predicted NMR chemical shifts with the NMR chemical shifts contained in the COLMAR database.[17−19] For a total of 179 $^{13}C-^{1}H$ moieties, the average prediction errors for carbon and proton chemical shifts are 2.903 and 0.292 ppm. The comparison between predicted and experimental chemical shifts is shown in Figure S2.

Weighted Matching between Experimental and Predicted NMR Spectra

After 2D $^{13}C-^{1}H$ HSQC NMR spectra were predicted for all chemical compound candidates, the weighted matching algorithm by Munkres was applied to match the 2D $^{13}C-^{1}H$ HSQC spectra extracted for individual mixture compounds with the predicted 2D $^{13}C-^{1}H$ HSQC spectra.[34] The use of a weighted matching algorithm is motivated by the goal to find the closest matching peak pairs for each experimental spin system to each predicted spin system, provided that the total number of spins is the same. The matching results are ranked according to the chemical shift root−mean−square deviation (RMSD) (eq 1) between the experimental and predicted chemical shifts:

$$\mathrm{RMSD} = \left\{\sum_{i=1}^{N}\left[(C_{i,\mathrm{exp}} - C_{i,\mathrm{pred}})^2 + ((H_{i,\mathrm{exp}} - H_{i,\mathrm{pred}}) \times 10)^2\right]/2N\right\}^{1/2} \quad (1)$$

in which X_{exp} are the experimental chemical shifts, X_{pred} are the predicted chemical shifts, and N is the number of peaks in the spin system. A scaling factor of 10 is used to normalize the effects of ^{13}C and ^{1}H chemical shifts on the overall RMSD by correcting for the different chemical shift ranges of these nuclei.

DOI: 10.1021/acs.jproteome.7b00457
J. Proteome Res. 2017, 16, 3774−3786

Table 1 shows the matching result for valine in the 25-compound model mixture.

Table 1. Example of Chemical Shift (c.s.) Matching Results for Valine

functional group	predicted ^{1}H c.s. (ppm)[a]	predicted ^{13}C c.s. (ppm)[a]	expt. ^{1}H c.s. (ppm)	expt. ^{13}C c.s. (ppm)
$-C_{\gamma}H_3$	0.910	19.32	1.034	20.56
$-C_{\gamma}H_3$	0.960	19.32	0.983	19.23
$-C_{\beta}H_2$	2.220	31.07	2.267	31.69
$-C_{\alpha}H$	3.440	61.90	3.597	63.00
RMSD (ppm)	0.93			

[a]Empirical chemical shift prediction was obtained by use of the NMR predictor by Modgraph embedded in the MestReNova software.

Molecular Structure Motif Identification of Compounds

After matching and rank-ordering, predicted NMR spectra for a potentially large number of candidate compounds derived from FT-ICR MS with an experimentally extracted spin system generally yielded a large number of hits with a reasonably low chemical shift RMSD cutoff (<5 ppm). Although candidate compounds with lower RMSDs generally are more likely to be the true compound, it cannot be excluded that the true compound has a lower rank due to the limited NMR prediction accuracy or molecular structure degeneracy. Therefore, to simplify and speed up the identification of the true compound among the hundreds or even thousands of hits, we propose the following approach referred to as "molecular structural motif identification of chemical compounds" or MSMIC. Because the experimentally extracted spin system corresponds to a structural motif consisting only of carbons and protons of the true compound, the goal of MSMIC is to find all possible molecular structural motifs that correspond to the experimentally extracted spin system among all of the compound candidates. The common molecular structural motif among different compounds will generate similar chemical shifts because additional atoms and functional groups typically have only little influence on the NMR chemical shift prediction of spins that are part of the molecular structure motif.

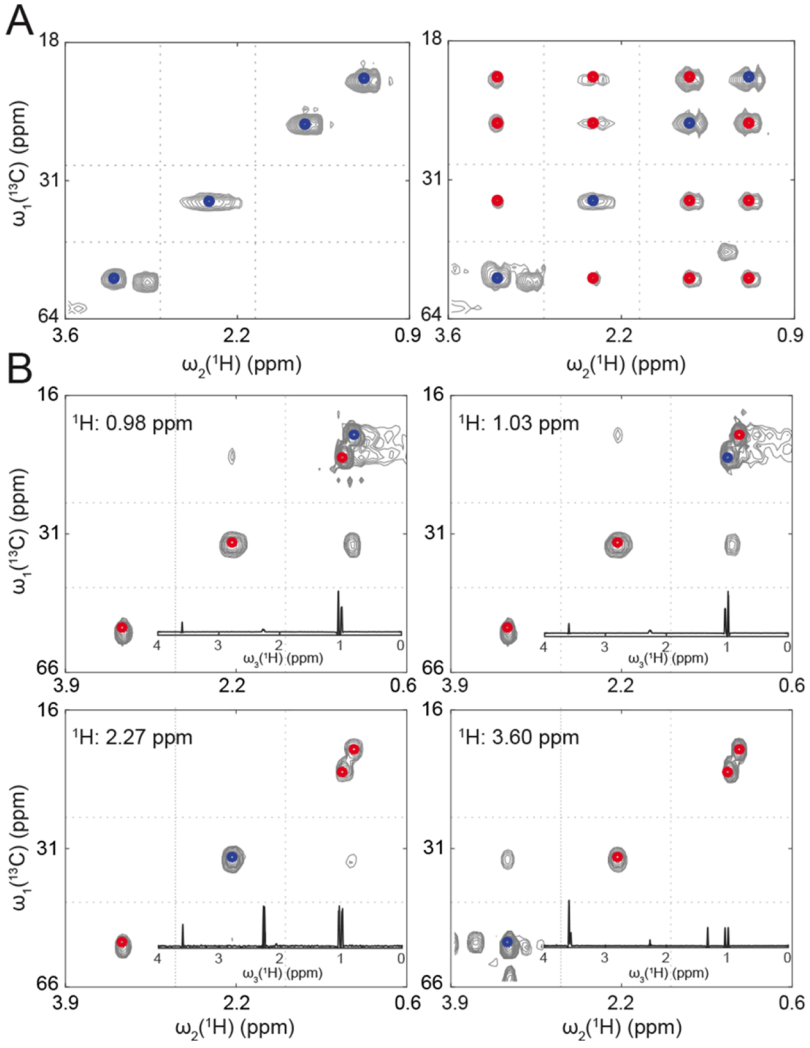


Figure 2. Putatively annotated valine spin system in a 25-metabolite model mixture extracted from 3D HSQC-TOCSY spectrum and confirmed by 2D TOCSY and 2D HSQC-TOCSY. Panel A: four single bond C−H cross-peaks (blue) of valine in the 2D HSQC (left) and 2D HSQC-TOCSY (right) spectra. The expected relay HSQC-TOCSY cross-peaks of the spin system are highlighted in red. Panel B: four different $\omega_1-\omega_2$ planes of the 3D HSQC-TOCSY spectrum belonging to valine. Blue peaks obey $\omega_2 = \omega_3$, and the red peaks are the other expected 3D cross-peaks of the valine spin system.

DOI: 10.1021/acs.jproteome.7b00457
J. Proteome Res. 2017, 16, 3774−3786

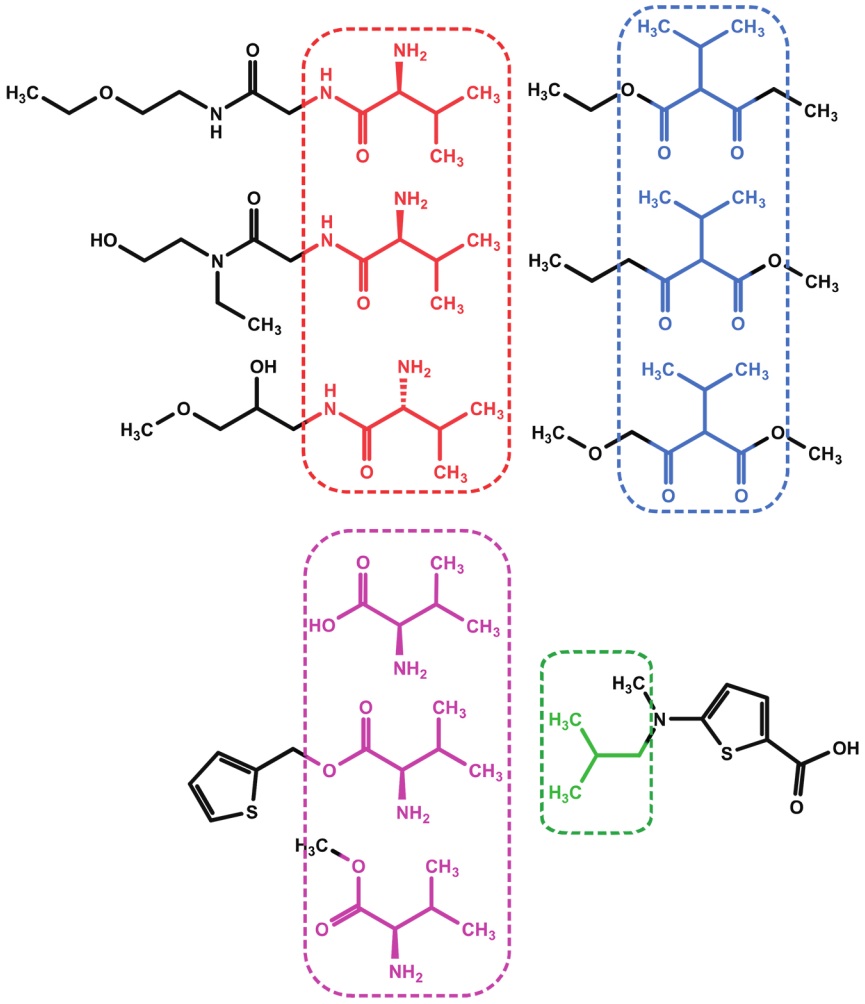


Figure 3. Molecular structural motifs identified by SUMMIT from 122 different compound candidates that all match the spin system of valine. The hits were sorted into different groups according to their common molecular motif that represents the NMR-derived spin system.

For instance, L-glutamine and glutathione share the common molecular structural motif ($HOOCCH(NH_2)CH_2CH_2CO$-$NH-$) and hence have similar chemical shifts for this motif (Figure S3). All hits (compound candidates) are sorted into groups according to their MSMICs by use of the nearest neighbor heavy atom for discrimination between different MSMICs. In a next step, molecular representatives of all high scoring MSMICs are selected for NMR experiments or used for quantum chemical calculations of their chemical shifts for the more accurate ranking of MSMICs. The best matching molecules are then either purchased or synthesized for NMR spiking experiments. This approach was implemented by use of in-house python scripts. Examples of MSMICs will be discussed below. In chemometrics, the maximum common substructure (MCS) approach is very efficient in identifying local structural similarities among large structural manifolds (>750 000).[35,36] Unfortunately, MCS is not able to identify the common motif that corresponds to a given spin system because it is not based on substructures connected by scalar J-couplings.

RESULTS AND DISCUSSION

Demonstration of SUMMIT MS/NMR to Identify Metabolites

First, we illustrate SUMMIT MS/NMR for the identification of metabolites based on the example of a spin system extracted from the 3D HSQC-TOCSY NMR spectrum of the 25-compound model mixture. The spin system with chemical shifts (δ_H, δ_C) of (3.597, 63.000), (2.267, 31.690), (1.034, 20.560), and (0.983, 19.230) ppm was used to match the predicted 2D HSQC NMR spectra of 57 881 compounds derived from elemental compositions corresponding to all above-threshold FT-ICR mass spectral peaks (see below). The spin−spin connectivity information is manifested in both 2D HSQC-TOCSY and 3D HSQC-TOCSY, which confirms that the four peaks indeed belong to the same spin system (Figure 2). The experimental chemical shifts were matched against the predicted chemical shifts of the 57 881 compounds by use of the weighted matching algorithm. On the basis of an RMSD cutoff of 5.0 ppm, 122 RMSD rank-ordered hits were returned. The top hit was valine with an RMSD of 0.93 ppm. Because valine was known to be one of the 25 compounds in the model mixture (which was independently confirmed by querying the chemical shifts of this spin system against the COLMAR $^1H(^{13}C)$-TOCCATA database,[18] yielding a low RMSD of 0.08 ppm for the database entry of valine), SUMMIT MS/NMR was successful in identifying (and verifying) this mixture component without depending on either a spectral NMR or MS database.

How should one proceed if the true compound does not exist in an NMR metabolomics database for validation, and how can one verify the true compound among dozens or hundreds of candidates returned by SUMMIT MS/NMR? Here, the

DOI: 10.1021/acs.jproteome.7b00457
J. Proteome Res. 2017, 16, 3774−3786

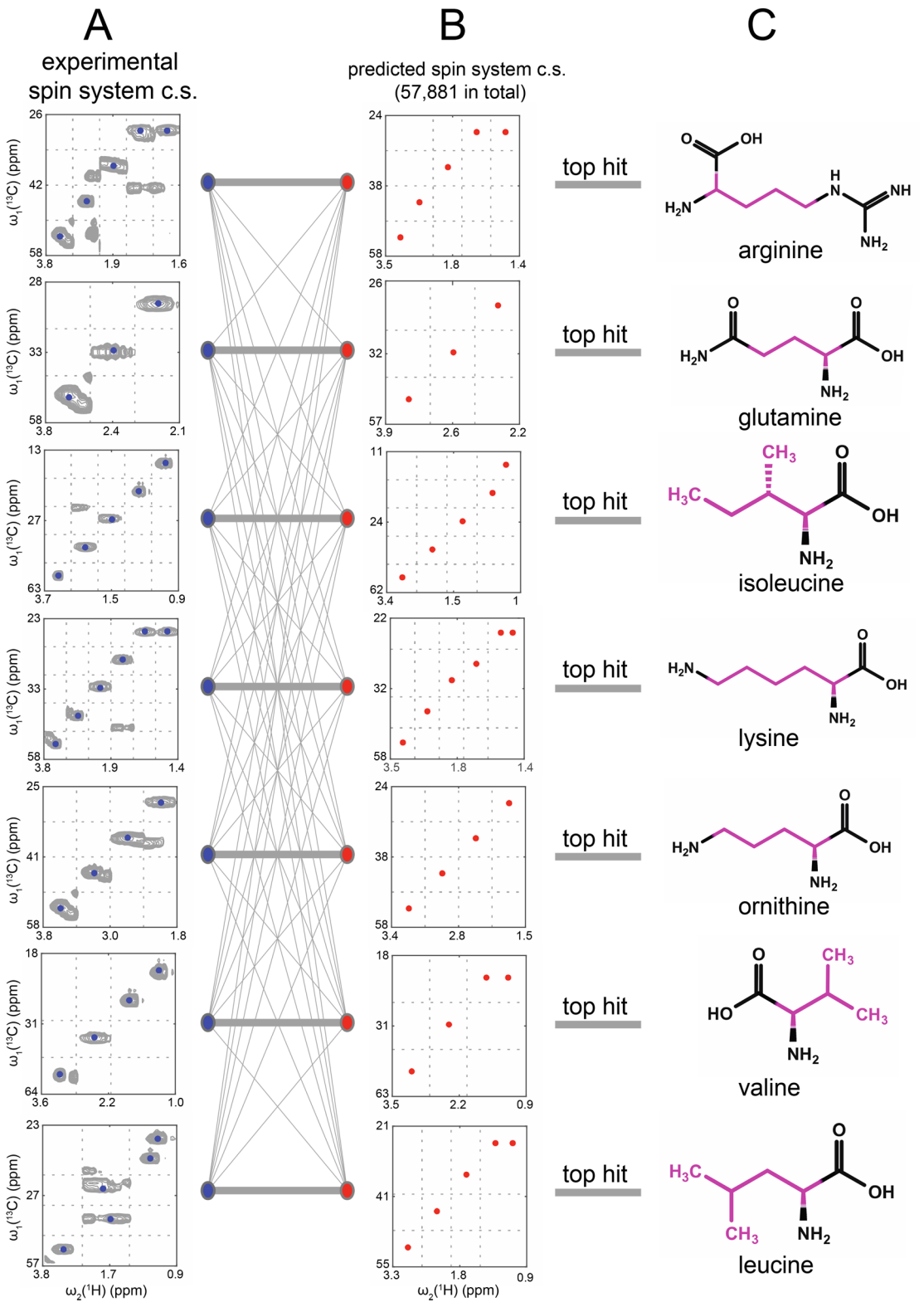


Figure 4. Identification of best matching metabolites in a 25-compound model mixture by SUMMIT MS/NMR. (A) Experimental 2D HSQC NMR spectra of metabolites extracted from 3D HSQC-TOCSY. Each spectrum is a collection of enlarged spectral regions (separated by dotted lines) that contain the corresponding cross-peaks. (B) Predicted 2D HSQC NMR spectra from FT-ICR MS-derived molecular structures (57 881 in total). Each experimental HSQC spectrum was compared with all 57 881 predicted HSQC spectra by maximum weighted bipartite matching. All returned hits were ranked according to their chemical shift RMSD. (C) Top hit compounds that belong to true compounds in the model mixture. Molecular substructures highlighted in magenta correspond to the molecular structural motifs (MSMIC) of the matched spin system. The spectra of panel B were sorted so that each experimental spectrum of panel A is adjacent to its top hit in panel B. To each edge of the graph connecting panels A and B belongs a chemical shift RMSD.

MSMIC approach described above comes to fruition. By first identifying the molecular structural motif of the true compound, it helps elucidate the complete structure of the true compound in a second step. Again, the spin system that

DOI: 10.1021/acs.jproteome.7b00457
J. Proteome Res. 2017, 16, 3774−3786

was eventually identified as valine is used as an example to demonstrate the strategy. For an unknown spin system with hundreds of compound candidates, the hits with lower RMSDs are more likely to correspond to the true compound. To verify the identity of the true compound beyond the limited NMR database information, quantum-chemical calculations followed by a NMR spiking experiment was adopted as the "gold standard" for compound verification.[37] In fact, before proceeding to verify the true compound by labor-intensive NMR spiking experiments for the 122 hits, the molecular structural motifs that reflect the possible substructures of the unknown spin systems are extracted, which both simplifies and speeds up the verification process. Examples of molecular structural motifs identified among the 122 hits are shown in Figure 3. As compounds with a common motif are expected to have similar chemical shifts, the next step is to select one or two compounds with low RMSD by quantum-chemical calculation in each cluster and perform NMR spiking experiments. Hence, the 122 initial hits were further reduced to fewer than 10 compound candidates for the verification of the molecular structural motif of the true compound. After confirmation of the MSMIC that belongs to an unknown spin system, further validation steps are performed to verify the entire compound as explained below.

Application to a 25-Compound Model Mixture

NMR and FT-ICR MS Data-Derived Information. On the basis of the maximal clique approach to automatically extract spin systems (see Methods section), 49 spin systems were extracted from the 3D $^{13}C-^{1}H$ HSQC-TOCSY NMR spectrum of the 25-compound model mixture. All extracted spin systems included two or more spins (one-spin systems were not included, see Methods section). Twenty-six spin systems were identified by SUMMIT MS/NMR and verified based on COLMAR $^{1}H(^{13}C)$-TOCCATA database query; two unknown spin systems could not be annotated and were classified as false positives because each resonance in these spin systems belongs to other spin systems as determined by visual inspection of the 2D TOCSY and HSQC-TOCSY spectra. Twenty-one spin systems were identified as partially or fully overlapping spin systems after spin system refinement as described in the Methods section. Eighty neutral molecular formulas (rms mass error 0.07 ppm) were obtained from the 100 highest magnitude FT-ICR mass spectral peaks by identifying elemental compositions matched to the METLIN database (see above) with <0.15 ppm mass error threshold (Figure S4).[32] (Peaks not originating from the 25 metabolites likely belong to impurities in the purchased compounds.) For each mass peak, $[M + H]^+$, $[M + Na]^+$, $[M+K]^+$, $[M+ACN+H]^+$, $[M+ACN+Na]^+$, and $[M +2Na-H]^+$ (in which M is the metabolite or its derivative) were considered as possible adducts. There were 57 881 molecular structures corresponding to the 80 molecular formulas according to the ChemSpider database (presently containing over 58 million molecular structures).[31] By comparison, if the mass error threshold was set to 1.0 ppm, 92 distinct molecular formulas were obtained with rms mass error 0.17 ppm, corresponding to 68 173 structures. In mixture analysis by MS, it is possible that intra- and interdimers and multimers may be generated. For example, ESI MS can yield both $[M + H]^+$ and $[2M+2H]^{2+}$ ions, which have the same monoisotopic mass-to-charge ratio, that is, $[^{12}C_c\,^{1}H_{(h+1)}\,^{14}N_n\,^{16}O_o\,^{31}P_p\,^{32}S_s]^+$ and $[^{12}C_{2c}\,^{1}H_{(2h+2)}\,^{14}N_{2n}\,^{16}O_{2o}\,^{31}P_{2p}\,^{32}S_{2s}]^{2+}$. However, the dimer is readily recognized by an m/z separation of 0.5 between $^{12}C_{2c}{}^{2+}$ and its $[^{12}C_{(2c-1)}\,^{13}C_1]^{2+}$ isotope peak. We did not observe any multimers of the reported metabolites.

Identification of Metabolites in the 25-Compound Model Mixture. After application of the weighted matching algorithm to match the 26 spin systems with the predicted 2D NMR HSQC spectra, the hits for each spin system were sorted according to the best matching chemical shift RMSD. Figure 4 shows the weighted matching scheme for the identification of metabolites by SUMMIT MS/NMR. Seven mixture compounds were returned as the top hits, namely valine, lysine, glutamine, isoleucine, arginine, and ornithine, and four compounds ranked between 2 and rank 10 [including fructose (ranked 4) of a total of 6 carbohydrates]. An additional six compounds ranked between the top 11−50 hits including adenosine (*vide infra*). Histidine and sucrose ranked 52th and 61th. Shikimic acid was not returned in the top 100 hits. However, a compound that has the same MCMIC as shikimic acid is the top hit. Four of the six carbohydrates were not in the top 100 hits due to high structural degeneracy and limited NMR chemical shift prediction accuracy. Finally, although the molecular weight of alanine (89.09320 Da) falls below the low-m/z limit of the FT-ICR MS, it can easily be identified by low resolution MS (e.g., quadrupole ion trap). Table 2 shows the matching results for the 25-compound model mixture.

Table 2. SUMMIT Results for 25-Compound Model Mixture (With Mass Error Cutoff of 0.15 ppm)

compound	rank	total hits (c.s. RMSD < 5.0 ppm)	percentile = rank/total number of hits	c.s. RMSD (ppm)	mass error (ppm)
valine	1	122	0.8%	0.93	−0.04
lysine	1	39	2.6%	1.54	0.04
glutamine	1	140	0.7%	1.36	0.07
isoleucine	1	16	6.3%	1.77	0
arginine	1	81	1.2%	2.02	0.10
ornithine	1	62	1.6%	2.05	0.05
leucine	1	13	7.7%	2.06	0
threonine	3	79	3.8%	1.71	−0.08
fructose	4	116	3.4%	1.32	0.05
carnitine	4	22	18.2%	3.32	0.12
cysteine	10	142	7.0%	2.17	−0.05
inosine	13	99	13.1%	2.27	−0.02
citrulline	23	65	35.4%	2.76	0.12
methionine	29	133	21.8%	2.53	0.03
serine	30	514	5.8%	2.2	0.07
proline	30	91	33.0%	2.55	0
adenosine	44	92	47.8%	2.74	0
histidine	52	181	28.7%	2.56	0.11
sucrose	61	117	52.1%	2.06	0.04

Table 3 shows the effect of (+) ESI FT-ICR mass spectral peak magnitude threshold on the number of possible ChemSpider structures and hit rank for the 25 metabolites mixture. For example, the number of possible structures drops from 57 881 to 16 030 or 9162 for a MS peak magnitude threshold increase from the top 100 highest-magnitude peaks to just the top 30 or 20 highest-magnitude peaks.

Although adenosine in the 25-compound mixture was low-ranked (rank 44 among 92 hits), it shares the ribose ring as the same common molecular structural motif with all other top 43 hits. Shikimic acid, galactose, glucose, lactose, and ribose were not identified (because chemical shift RMSDs > 5.0 ppm), but their molecular structural motifs were correctly identified by

DOI: 10.1021/acs.jproteome.7b00457
J. Proteome Res. 2017, 16, 3774−3786

Table 3. Effect of Cutoff Threshold of Mass Peak Amplitudes on the Rank of SUMMIT Results for 25-Compound Model Mixture (0.15 ppm Mass Error Cutoff)

index	compound	rank: top 30 mass peaks	rank: top 20 mass peaks	rank: all mass peaks	c.s. RMSD (ppm)
1	valine	1	1	1	0.83
2	lysine	1	1	1	1.44
3	glutamine	1	1	1	1.35
4	isoleucine	1	1	1	1.86
5	arginine	1	1	1	1.95
6	ornithine	1	1	1	1.92
7	leucine	1	1	1	1.84
8	threonine	2	N/A	3	1.55
9	fructose	4	4	4	1.36
10	carnitine	4	3	4	3.14
11	methionine	6	3	29	2.18
12	proline	7	5	30	2.54
13	cysteine	7	N/A	10	1.93
14	citrulline	11	10	23	2.40
15	inosine	13	13	13	2.26
16	serine	17	11	30	2.07
17	histidine	18	9	52	2.58
18	adenosine	39	38	44	2.63
19	sucrose	61	61	61	1.77
structure manifolds		16 030	9162	57 881	

SUMMIT MS/NMR. These results show the ability of this approach to correctly identify structural motifs belonging to spin systems identified by NMR. The remaining ambiguity among molecules with the correct structural motif is due to limitations of the empirical chemical shift predictor, which can be alleviated in part by applying more accurate, but also more expensive, quantum-chemistry based chemical shift calculations to the top hits.[38]

To reduce the number of false-positive compounds returned by weighted matching, we optimized the mass error threshold for molecular formulas determination to 0.15 ppm, which significantly reduced the number of false compounds and improved the rankings of the true compounds. As an example, mixture compound leucine ranked seventh of a total of 29 hits if a 1.0 ppm mass error threshold was used, but it ranked as the top hit of 13 returned hits after lowering the mass error threshold to 0.15 ppm. Note that the average mass error for the true 25 compounds is 30 ppb and the maximum mass error is less than 120 ppb, which demonstrates the high mass accuracy that can be achieved from an ultrahigh-resolution FT-ICR mass spectrum for a metabolic mixture. Therefore, for an unknown metabolite in a metabolomics mixture, focusing on the low ppm mass accuracy molecular formulas will lead to fewer molecular formulas and thereby facilitate the true compound identification. It should be noted that the low ppm mass accuracy cutoff varies from sample to sample, but it can be determined for each sample by the identification of known abundant metabolites in the mixture.

Validation of Putative Metabolite Structure by FT-ICR MS/MS. MS/MS can serve to validate the top-ranked structures. As a demonstration, FT-ICR infrared multiphoton dissociation (IRMPD) was performed for the isolated precursor ions corresponding to SUMMIT MS/NMR top-ranked glutamine, lysine, arginine, and ornithine (Figure 5, Figures S5 and S6). Glutamine and lysine MS/MS yielded mass differences

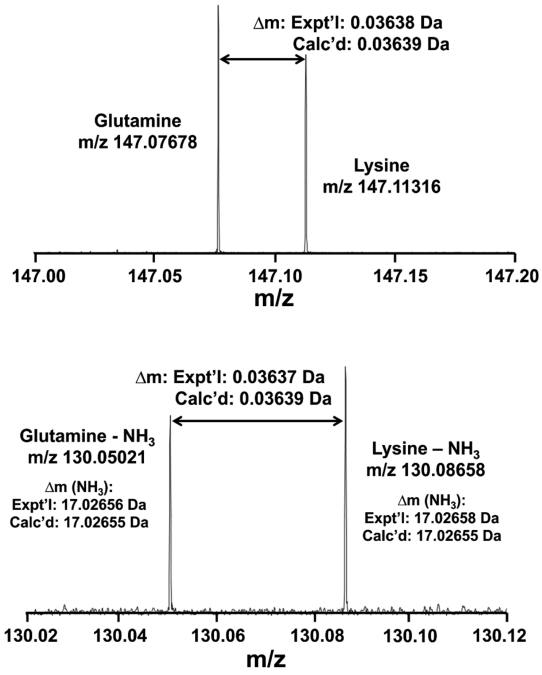


Figure 5. FT-ICR MS/MS of glutamine and lysine in 25 metabolite mixture. Glutamine and lysine MS/MS yields mass differences (between the precursor ion and product ion) of 17.02656 and 17.02658 Da (i.e., 0.01 mDa and 0.03 mDa deviation from the calculated mass 17.02655 Da) corresponding to loss of ammonia.

(between the precursor ion and product ion) of 17.02656 and 17.02658 Da (i.e., 0.01 mDa and 0.03 mDa deviation from the calculated mass 17.02655 Da) corresponding to loss of ammonia. Arginine and ornithine MS/MS yielded loss of ammonia (0.06 mDa and 0.05 mDa deviation from the calculated mass of 17.02655 Da) and loss of water (0.06 mDa and 0.06 mDa deviation from the calculated mass of 18.01056 Da). Collision-induced dissociation (CID) in a linear quadrupole ion trap yielded loss of ammonia, water, and carbon monoxide from valine precursor ion (Figure S7). Therefore, the product ion mass spectrum further supports the highest ranked SUMMIT-based structures. Although the information content of MS/MS fragment analysis varies from metabolite to metabolite, MS/MS is expected to be most helpful for SUMMIT MS/MS/NMR for the identification of larger molecules, such as secondary metabolites.

Application to *E. coli* Cell Lysate

NMR and FT-ICR MS Data-Derived Information. Three-hundred ninety-seven potential spin systems were extracted from the 3D HSQC-TOCSY NMR spectrum of the *E. coli* cell lysate by applying the maximal clique approach in full analogy to the model mixture. All extracted spin systems included three or more spins. Besides one-spin systems, two-spin systems were also excluded to avoid the generation of a potentially large number of false positives. We obtained a total of 1095 molecular formulas by searching the FT-ICR broadband accurate masses against the METLIN database (see above), leading to the generation of 914 947 candidate molecular structures by screening the ChemSpider database.

Identification of Known Metabolites in *E. coli*. In the 25-compound model mixture, all of the compounds are known and they are contained in the NMR database, thereby enabling testing of the SUMMIT MS/NMR method. Here, we first

DOI: 10.1021/acs.jproteome.7b00457
J. Proteome Res. 2017, 16, 3774−3786

apply SUMMIT MS/NMR to identify known metabolites in *E. coli* to test the power and limitations of the method by comparing the results with those obtained by querying the spectra directly against the COLMAR web server. The recently developed COLMARm web server module provides simultaneous analysis of 2D HSQC, 2D TOCSY, and 2D HSQC-TOCSY NMR spectra and is used to identify metabolites. Metabolites were first identified by querying the 2D HSQC against the COLMAR database and subsequently verified by 2D TOCSY and 2D HSQC-TOCSY by use of COLMARm. Forty-one metabolites could be identified with high confidence by COLMARm (2D HSQC cross-peak matching ratio >0.8 and more than 50% spin−spin connectivities showing up in the 2D TOCSY and 2D HSQC-TOCSY spectra), which are listed in the Supporting Information Table S1. The 41 metabolites were treated as "putatively annotated metabolites" to be verified by SUMMIT MS/NMR. When implementing SUMMIT MS/NMR to verify the metabolites identified by COLMAR, we compared the identified metabolites with the matching results returned for each extracted spin system. Verification results are reported in Supporting Information Table S2 for metabolites that fulfill the following conditions: they are ranked among the top 200 hits if the total number of hits with a chemical shift RMSD < 5.0 ppm was 400 or less or they are ranked in the top 50% percentile if the total number of hits with RMSD < 5.0 ppm exceeded 400. These criteria ensure that the most likely candidates are retained without making the pool unrealistically large. On the basis of cross-platform analytical methods to verify compounds, the identification and verification results by COLMAR and SUMMIT MS/NMR achieved level 2 confidence according to the Metabolomics Standards Initiative.[37]

The following 13 known metabolites were successfully verified by SUMMIT MS/NMR: L-glutamine, L-valine, maltose, cellobiose, N-acetyl-putrescine, L-glutamic-acid, D-glucose, spermidine, L-phenylalanine, L-tyrosine, N-α-acetyl-L-lysine, L-glutathione-reduced, and L-methionine. Adenosine, inosine, L-proline, leucine, pyridoxamine-5-phosphate-1, and guanosine could not be verified because not all of their cross-peaks showed up due to the relatively low abundance of these metabolites and the limited sensitivity of HSQC-TOCSY. However, by manually checking 2D HSQC-TOCSY and 3D HSQC-TOCSY, partial spin systems of these metabolites (covering 50% or more of the expected cross-peaks) could be identified. When implementing SUMMIT MS/NMR, we set the matching ratio cutoff to 1 to increase the identification accuracy when matching with FT-ICR MS-derived NMR spectra. For instance, if a compound contains a five-spin system, but only a four-spin subsystem could be reconstructed from 3D HSQC-TOCSY (e.g., because a resonance is very weak), the true (5-spin system) compound would not be returned as a hit by matching the experimental four-spin system with the FT-ICR MS-derived NMR spectra.

Therefore, SUMMIT MS/NMR will verify only the metabolites that are detectable by both analytical methods, providing high validation confidence across platforms. In addition, off-line LC fractionation can be applied prior to MS/NMR analysis to decrease the complexity of the metabolites mixture and increase the chance to identify more common metabolites by MS and NMR. Those metabolites that are identified by only one of the two analytical methods need to be further validated by other analytical methods, for example, HPLC-MS. In any case, HPLC retention time (especially when calibrated by spiking with the putative metabolite) can help further validate any metabolite assignments based on NMR, MS, or a combination of the two. Finally, metabolites that are not detected by positive ESI can often be detected by negative ESI. For example, the MS1 accurate masses for acetyl-L-glutamine, DL-α-glycerol-phosphoric acid, D-glucuronic acid, methyl-uridine, deoxythymidine monophosphate, uridine monophosphate, and cystathionine were detected by use of negative ESI (rms mass error 0.18 ppm), and those identities were confirmed by NMR.

Identification of Unknown Metabolites in *E. coli* by SUMMIT MS/NMR. The primary aim of SUMMIT MS/NMR is to identify unknown compounds that are not catalogued in current NMR and MS metabolomics databases. Unknown spin systems show self-consistent spin−spin connectivities in the 3D HSQC-TOCSY spectrum, but do not match any compound in the NMR database. Here, SUMMIT MS/NMR identified up to 15 unknown spin systems (compounds) in *E. coli* cell lysate. For instance, the spin system with chemical shifts (δ_H, δ_C) identified as (1.167,19.542), (3.618,59.207), (3.632,75.944), (3.767,73.296), (4.032,70.725), and (5.573,98.048) ppm shows high confidence as a true positive spin system (Figure 6), but it could not be assigned to any known compound after querying against the COLMAR database. However, after matching against 914 947 predicted NMR spectra, 12 hits (RMSD < 5.0 ppm) were returned, and four molecular structural motifs were identified (Figure S8). As a proof of principle for the identification of true compounds, we applied quantum-chemical calculations to two selected compounds among the 12 hits, namely L-fucosamine and 6-desoxy-D-glucosamine. 6-Desoxy-D-glucosamine was the top hit among the 12 hits. L-Fucosamine has been found as a constituent of mucopolysaccharides of certain enteric bacteria (e.g., *Citrobacter fiemdii*), but the existence of L-fucosamine in *E. coli* was previously unknown.[39] Quantum-chemical calculations of NMR chemical shifts for these two compounds return a lower RMSD for L-fucosamine, and hence, L-fucosamine is more likely to be the true compound than 6-desoxy-D-glucosamine, consistent with the literature (Table S3). Although the true identity of this spin system remains uncertain, SUMMIT MS/NMR provides a small list of likely candidates, which represents actionable information for the identification of the true compound.

Pyroglutamic acid, which at the outset of this study was not part of the COLMAR $^{1}H(^{13}C)$-TOCCATA database, represents another instructive example of the SUMMIT approach. SUMMIT MS/NMR successfully extracted the spin system and returned pyroglutamic acid as the 116th hit (Figure S9). Independently and at about the same time, the COLMAR $^{1}H(^{13}C)$-TOCCATA database increased by 284 compounds,[17] including pyroglutamic acid, thereby enabling the identification of pyroglutamic acid as a known metabolite, confirming the SUMMIT results. For the *E. coli* cell lysate, SUMMIT returned 15 unknown spin systems along with their candidate compounds. To maximize the confidence of the unknown spin systems, we included all of the pairwise-connected spins that appear along the 1D ω_3 (^{1}H) trace, and none of the peaks in the spin system matched the NMR database. The unknown spin systems and compound candidates (top hit) are listed in the Supporting Information Table S4.

Finally, we note that some unknown spin systems have multiple candidate compounds, whereas others do not match any candidate compounds based on our metrics. Extending the mass range of FT-ICR MS will be helpful to incorporate all possible compounds and find the compound candidates for these

DOI: 10.1021/acs.jproteome.7b00457
J. Proteome Res. 2017, 16, 3774−3786

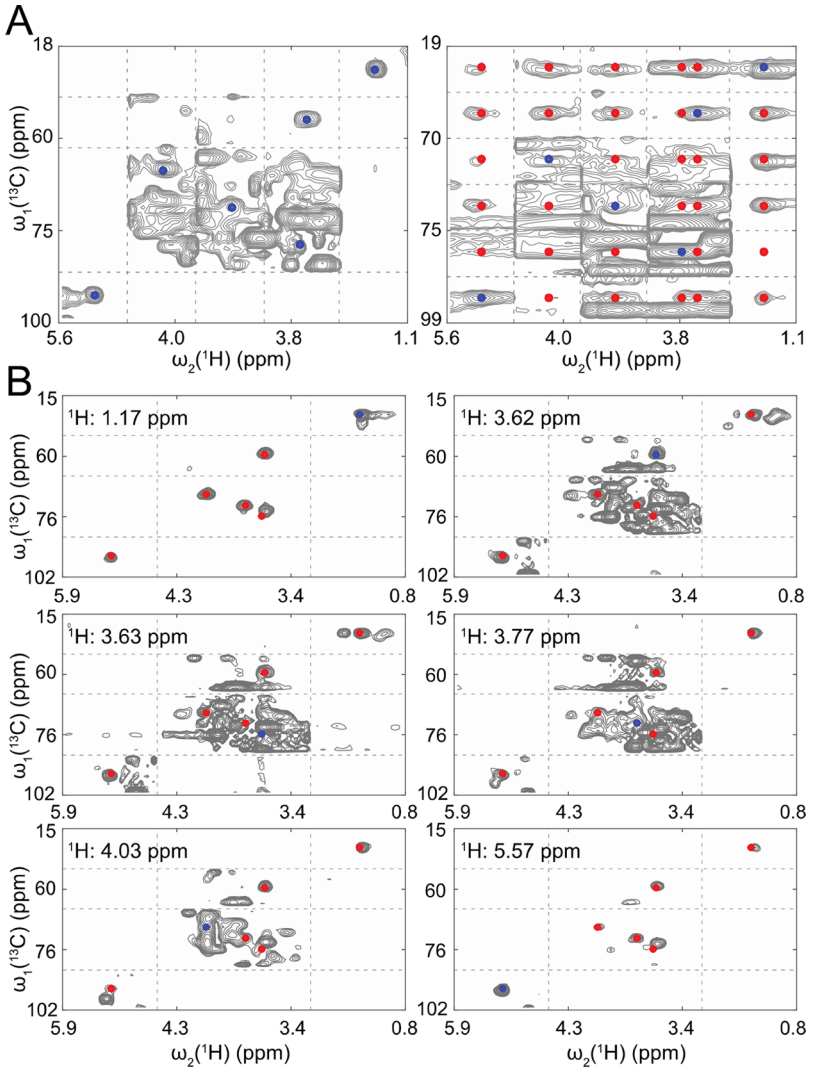


Figure 6. Spin system of an unknown compound from an *E. coli* cell lysate extracted from 3D HSQC-TOCSY and verified by 2D TOCSY and 2D HSQC-TOCSY (2D TOCSY is not shown). (A) Cross-peaks of the unknown compound shown in 2D HSQC (left, blue cross-peaks) and 2D HSQC-TOCSY (right, blue and red cross-peaks) spectra. (B) Six cross-peaks of the unknown compound depicted in six different 2D slices (ω_1,ω_2) at fixed ω_3 frequency of the 3D HSQC-TOCSY spectrum (blue symbols, diagonal peaks; red symbols, cross-peaks).

unknown spin systems. Nevertheless, SUMMIT MS/NMR provides powerful fingerprints, based on spin system information, molecular formulas, and compound candidates in complex biological mixtures, thereby greatly assisting the analysis of complex metabolomics mixtures whose compositions are only partially known, without being limited to spectroscopic databases. SUMMIT is expected to find fruitful applications to support key objectives of contemporary metabolomics research, including the discovery of new biochemical pathways and biomarkers.

EXPERIMENTAL SECTION

Sample Preparation

A 25-compound metabolite mixture contained adenosine, alanine, arginine, carnitine, citrulline, cysteine, fructose, galactose, glucose, glutamine, histidine, inosine, isoleucine, lactose, leucine, lysine, methionine, ornithine, proline, ribose, serine, shikimate, sucrose, threonine, and valine. For the NMR experiments, the final concentration of each metabolite was 1 mM in 600 μL of D_2O with 20 mM phosphate buffer and 0.1 mM DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid) for chemical shift referencing. The same 25 compounds were used for the MS sample, which was prepared in 50%/50% (v/v) ACN/H_2O solution with 0.1% formic acid. The final concentration of each metabolite for MS was 10 μM. All chemicals and solvents were obtained from Sigma-Aldrich and Fisher Scientific Corporation.

E. coli BL21(DE3) cells were cultured at 37 °C with shaking at 250 rpm in M9 minimum medium with glucose (natural abundance, 5 g/L) added as the sole carbon source. One liter of culture at OD 1 was centrifuged at 5000 × *g* for 20 min at 4 °C, and the cell pellet was resuspended in 50 mL of 50 mM phosphate buffer at pH 7.0. The cell suspension was then subjected to centrifugation for cell pellet collection. The cell pellet was resuspended in 10 mL of ice-cold water and freeze−thawed three times. The sample was centrifuged at 20 000 × *g* at 4 °C for 15 min to remove cell debris. Prechilled methanol and chloroform were sequentially added to the supernatant under vigorous vortexing at an H_2O/methanol/chloroform ratio of 1:1:1 (v/v/v). The mixture was then left at −20 °C overnight for phase separation. Next, it was centrifuged at

DOI: 10.1021/acs.jproteome.7b00457
J. Proteome Res. 2017, 16, 3774−3786

4000 × *g* for 20 min at 4 °C, and the clear upper hydrophilic phase was collected and subjected to rotary evaporation to reduce the methanol content. Finally, the sample was lyophilized. The dry sample was then divided into two parts: one for MS and one for NMR analysis. The NMR sample was prepared by dissolving the material in 200 μL of D_2O with 20 mM phosphate buffer and 0.1 mM DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid) for chemical shift referencing, and then transferred to a 3 mm NMR tube. Then 1.5−2 mg of *E. coli* cell lysate was dissolved in 200 μL of H_2O, and 10 μL of that was diluted 10-fold with 50%/50% (v/v) ACN/H_2O with 0.1% formic acid. The resulting solution was centrifuged at 13 000 rpm at 4 °C for 5 min, and the supernatant was used for direct-infusion MS.

NMR Experiments and Data Processing

2D $^{13}C-^{1}H$ HSQC, 2D $^{1}H-^{1}H$ TOCSY, 2D $^{13}C-^{1}H$ HSQC-TOCSY, and 3D $^{13}C-^{1}H$ HSQC-TOCSY spectra of the 25-compound model mixture and *E. coli* cell lysate were collected. All NMR spectra of the 25-compound model mixture and *E. coli* cell lysate were acquired with a Bruker AVANCE solution-state NMR spectrometer equipped with a cryogenically cooled TCI probe at 850 MHz proton frequency at 298 K. The 2D $^{13}C-^{1}H$ HSQC spectra of the 25-compound model mixture and *E. coli* cell lysate were collected with 256 t_1 and 1024 t_2 complex points. The measurement time was ∼2 h. The spectral width along the indirect and the direct dimensions was 34205.6 and 10204.1 Hz. The number of acquisitions per t_1 increment was 8. The transmitter frequency offset was 80 ppm in the ^{13}C dimension and 4.7 ppm in the ^{1}H dimension. The 2D $^{1}H-^{1}H$ TOCSY spectra of the 25-compound model mixture and *E. coli* cell lysate were collected with 512 t_1 and 1024 t_2 complex points. The measurement time was ∼4 h. The spectral width along the indirect and direct dimensions was set to 10204.1 Hz. The number of acquisitions per t_1 increment was 8. The transmitter frequency offset was 4.7 ppm in both ^{1}H dimensions. The TOCSY mixing time was set to 120 ms after optimization of the isotropic mixing time (Figure S10). 2D $^{13}C-^{1}H$ HSQC-TOCSY spectra of the 25-compound model mixture and *E. coli* cell lysate were collected with 512 t_1 and 2048 t_2 complex points. The measurement time was ∼8.5 h. The spectral width along the indirect and the direct dimensions was 34205.6 and 10204.1 Hz. The TOCSY mixing time for 2D $^{13}C-^{1}H$ HSQC-TOCSY was set to 120 ms. The number of acquisitions per t_1 increment was 16. The transmitter frequency offset was 80 ppm in the ^{13}C dimension and 4.7 ppm in the ^{1}H dimension. 3D $^{13}C-^{1}H$ HSQC-TOCSY spectra of the 25-compound model mixture and *E. coli* cell lysate were collected with 64 t_1, 128 t_2, and 2048 t_3 complex points. The measurement time was ∼113 h. The spectral width along the indirect and the direct dimensions was 34205.6, 10204.1, and 10204.1 Hz. The number of scans per t_1 increment was 8. The transmitter frequency offset was 80 ppm in the ^{13}C dimension and 4.7 ppm in the ^{1}H dimension. The data were zero-filled two-fold along the ^{13}C dimension, Fourier transformed, and phase- and baseline-corrected by use of NMRPipe.[40] Sparky was used for peak-picking in all spectra.[41] All spectra were converted to MATLAB format for maximal clique analysis.

FT-ICR MS Experiments and Processing

A custom-built 9.4 T Fourier transform ion cyclotron resonance mass spectrometer was used for sample analysis.[21] A 25 metabolites mixture (10 μM) and *E. coli* extract sample (in 50% ACN, 50% water, and 0.1% formic acid) were ionized by positive or negative nanoelectrospray at a flow rate of 0.3 μL/min and accumulated in an external linear quadrupole ion trap. Ions were then transferred through an octopole ion guide to the ICR cell for broadband and tandem mass spectra acquisition. The transfer time was set to 0.35 ms for lower mass range analysis [m/z 107−270] and 0.55 ms for higher mass range analysis [m/z 265−400]. MS/MS fragmentations for glutamine, lysine, arginine, and ornithine were performed by infrared multiphoton dissociation (IRMPD), and precursor ions were isolated externally with a quadrupole mass filter and internally by stored waveform inverse Fourier transform (SWIFT) excitation in the ICR cell.[42] IRMPD was performed with a 40 W, 10.6 mm, CO_2 laser (Synrad, Mukilteo, WA, USA), fitted with a 2.5× beam-expander. The laser beam was directed to the center of the cell through an off-axis BaF_2 window. Photon irradiation was for 500 ms at 40−90% laser power (16−36 W). MS/MS fragmentation for valine was performed with a Velos Pro ion trap mass spectrometer with normalized collisional energy 22 (ThermoFisher parameter setting). Broadband and tandem mass spectra were acquired from m/z 107−2000, with a 6 s time-domain acquisition period. Five-hundred time-domain transients were digitized and signal-averaged. All data were stored as .DAT files. All time-domain data were Hanning apodized, zero-filled, and fast Fourier transformed to yield magnitude-mode mass spectra. Frequency-to-m/z conversion was performed with a two-term calibration equation.[43,44] Mass calibration was performed by dual spray spanning m/z 112−410. For positive ESI, the custom-prepared standard mix included cytosine, caffeine, biotin, adenosine, Val-Ala-Pro-Gly, and [des-Tyr1]-methionine enkephalinamide. For negative ESI, Agilent ESI-L Low Concentration Tuning Mix (Agilent, Santa Clara, CA) was used for calibration. After dual spray calibration, high magnitude peaks (6 peaks for ESI positive mode and 5 peaks for ESI negative mode) in m/z range 112−410 were chosen as internal standards and used for calibration during sample direct infusion into the mass spectrometer. Data were manually interpreted by use of Predator Analysis (version 4.1.9) software.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.7b00457.

Spin system refinement procedure; predicted chemical shifts compared with their experimental chemical shifts of 25-compound model mixture; chemical shifts in 2D $^{13}C-^{1}H$ HSQC spectra; FT-ICR MS spectrum of *E. coli* cell lysate; FT-ICR MS/MS spectrum of arginine; FT-ICR MS/MS spectrum of ornithine; MS/MS spectrum of valine; motif identification of compound candidates for the spin system of unknown compound; spin system of pyroglutamic acid spin system; effect of TOCSY mixing time on magnetization transfer efficiency of inosine; metabolites of *E. coli* cell lysate identified by COLMARm web server and database; metabolites identified in *E. coli* cell lysate and verified by COLMAR web server and SUMMIT MS/NMR; quantum-chemical calculation of NMR chemical shifts; unknown spin systems identified in *E. coli* cell lysate by 3D HSQC-TOCSY and compound candidates (PDF)

DOI: 10.1021/acs.jproteome.7b00457
J. Proteome Res. 2017, 16, 3774−3786

AUTHOR INFORMATION

Corresponding Authors

*E-mail: bruschweiler.1@osu.edu.
*E-mail: marshall@magnet.fsu.edu.

ORCID

Alan G. Marshall: 0000-0001-9375-2532
Rafael Brüschweiler: 0000-0003-3649-4543

Author Contributions

‖These authors contributed equally.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Felix Hoffmann for performing quantum-chemical chemical shift calculations for Table S3. This work was supported by the National Institutes of Health (Grant No. R01 GM 066041 and SECIM (Southeast Center for Integrated Metabolomics) Grant No. U24 DK097209-01A1 to R.B.). The FT-ICR experiments were performed at the National High Magnetic Field Laboratory, supported by the National Science Foundation (Grant No. DMR-11-57490). All NMR experiments were performed at the CCIC NMR facility at OSU.

REFERENCES

(1) Markley, J. L.; Brüschweiler, R.; Edison, A. S.; Eghbalnia, H. R.; Powers, R.; Raftery, D.; Wishart, D. S. The future of NMR-based metabolomics. *Curr. Opin. Biotechnol.* **2017**, *43*, 34−40.

(2) Huan, T.; Tang, C. Q.; Li, R. H.; Shi, Y.; Lin, G. H.; Li, L. MyCompoundID MS/MS search: Metabolite identification using a library of predicted fragment-ion-spectra of 383,830 possible human metabolites. *Anal. Chem.* **2015**, *87*, 10619−10626.

(3) Yilmaz, A.; Rudolph, H. L.; Hurst, J. J.; Wood, T. D. High-Throughput metabolic profiling of soybean leaves by fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **2016**, *88*, 1188−1194.

(4) Rathahao-Paris, E.; Alves, S.; Junot, C.; Tabet, J. C. High resolution mass spectrometry for structural identification of metabolites in metabolomics. *Metabolomics* **2016**, *12*, 1 DOI: 10.1007/s11306-015-0882-8.

(5) Cloarec, O.; Dumas, M. E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; Nicholson, J. Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic 1H NMR data sets. *Anal. Chem.* **2005**, *77*, 1282−1289.

(6) Hao, J.; Liebeke, M.; Sommer, U.; Viant, M. R.; Bundy, J. G.; Ebbels, T. M. D. statistical correlations between NMR spectroscopy and direct infusion FT-ICR mass spectrometry aid annotation of unknowns in metabolomics. *Anal. Chem.* **2016**, *88*, 2583−2589.

(7) Wei, S. W.; Zhang, J.; Liu, L. Y.; Ye, T.; Nagana Gowda, G. A.; Tayyari, F.; Raftery, D. Ratio analysis nuclear magnetic resonance spectroscopy for selective metabolite identification in complex samples. *Anal. Chem.* **2011**, *83*, 7616−7623.

(8) Gu, H. W.; Nagana Gowda, G. A.; Neto, F. C.; Opp, M. R.; Raftery, D. RAMSY: ratio analysis of mass spectrometry to improve compound identification. *Anal. Chem.* **2013**, *85*, 10771−10779.

(9) Crockford, D. J.; Holmes, E.; Lindon, J. C.; Plumb, R. S.; Zirah, S.; Bruce, S. J.; Rainville, P.; Stumpf, C. L.; Nicholson, J. K. Statistical heterospectroscopy, an approach to the integrated analysis of NMR and UPLC-MS data sets: application in metabonomic toxicology studies. *Anal. Chem.* **2006**, *78*, 363−371.

(10) Bingol, K.; Zhang, F.; Bruschweiler-Li, L.; Brüschweiler, R. Carbon backbone topology of the metabolome of a cell. *J. Am. Chem. Soc.* **2012**, *134*, 9006−11.

(11) Clendinen, C. S.; Pasquel, C.; Ajredini, R.; Edison, A. S. 13C NMR metabolomics: INADEQUATE network analysis. *Anal. Chem.* **2015**, *87*, 5698−5706.

(12) Komatsu, T.; Ohishi, R.; Shino, A.; Kikuchi, J. Structure and metabolic-flow analysis of molecular complexity in a 13C-labeled tree by 2D and 3D NMR. *Angew. Chem., Int. Ed.* **2016**, *55*, 6000−3.

(13) Bingol, K.; Brüschweiler, R. NMR/MS translator for the enhanced simultaneous analysis of metabolomics mixtures by NMR spectroscopy and mass spectrometry: application to human urine. *J. Proteome Res.* **2015**, *14*, 2642−2648.

(14) Walker, L. R.; Hoyt, D. W.; Walker, S. M., 2nd; Ward, J. K.; Nicora, C. D.; Bingol, K. Unambiguous metabolite identification in high-throughput metabolomics by hybrid 1D 1H NMR/ESI MS1 approach. *Magn. Reson. Chem.* **2016**, *54*, 998−1003.

(15) Bingol, K.; Bruschweiler-Li, L.; Yu, C.; Somogyi, A.; Zhang, F.; Brüschweiler, R. Metabolomics beyond spectroscopic databases: a combined MS/NMR strategy for the rapid identification of new metabolites in complex mixtures. *Anal. Chem.* **2015**, *87*, 3864−70.

(16) Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom. Rev.* **1998**, *17*, 1−35.

(17) Bingol, K.; Zhang, F. L.; Bruschweiler-Li, L.; Brüschweiler, R. TOCCATA: a customized carbon total correlation spectroscopy NMR metabolomics database. *Anal. Chem.* **2012**, *84*, 9395−9401.

(18) Bingol, K.; Bruschweiler-Li, L.; Li, D. W.; Brüschweiler, R. Customized metabolomics database for the analysis of NMR 1H-1H TOCSY and 13C-1H HSQC-TOCSY spectra of complex mixtures. *Anal. Chem.* **2014**, *86*, 5494−501.

(19) Bingol, K.; Li, D. W.; Bruschweiler-Li, L.; Cabrera, O. A.; Megraw, T.; Zhang, F. L.; Brüschweiler, R. Unified and isomer-specific NMR metabolomics database for the accurate analysis of 13C-1H HSQC spectra. *ACS Chem. Biol.* **2015**, *10*, 452−459.

(20) Bingol, K.; Li, D. W.; Zhang, B.; Brüschweiler, R. Comprehensive metabolite identification strategy using multiple two-dimensional NMR spectra of a complex mixture implemented in the COLMARm web server. *Anal. Chem.* **2016**, *88*, 12411−12418.

(21) Kaiser, N. K.; Quinn, J. P.; Blakney, G. T.; Hendrickson, C. L.; Marshall, A. G. A novel 9.4 T FTICR mass spectrometer with improved sensitivity, mass resolution, and mass range. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 1343−51.

(22) Kim, S.; Rodgers, R. P.; Marshall, A. G. Truly "exact" mass: Elemental composition can be determined uniquely from molecular mass measurement at similar to 0.1 mDa accuracy for molecules up to similar to 500 Da. *Int. J. Mass Spectrom.* **2006**, *251*, 260−265.

(23) Marshall, A. G.; K, L. C.; Enke, C. G.; Hendrickson, C. L. Dynamic range extension in FT-ICR mass spectrometry by spectral segmentation. *Proceedings of 63rd Am. Soc. for Mass Spectrometry Annual Conference on Mass Spectrometry and Related Topics*, St. Louis, MO, May 31−June 5, 2015, Poster TP 121.

(24) Hughey, C. A.; Hendrickson, C. L.; Rodgers, R. P.; Marshall, A. G.; Qian, K. Kendrick mass defect spectrum: a compact visual analysis for ultrahigh-resolution broadband mass spectra. *Anal. Chem.* **2001**, *73*, 4676−81.

(25) Wu, Z.; Rodgers, R. P.; Marshall, A. G. Two- and three-dimensional van krevelen diagrams: a graphical analysis complementary to the kendrick mass plot for sorting elemental compositions of complex organic mixtures based on ultrahigh-resolution broadband fourier transform ion cyclotron resonance mass measurements. *Anal. Chem.* **2004**, *76*, 2511−6.

(26) Marshall, A. G.; Rodgers, R. P. Petroleomics: chemistry of the underworld. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 18090−5.

(27) Misiak, M.; Kozminski, W. Determination of heteronuclear coupling constants from 3D HSQC-TOCSY experiment with optimized random sampling of evolution time space. *Magn. Reson. Chem.* **2009**, *47*, 205−209.

(28) Reardon, P. N.; Marean-Reardon, C. L.; Bukovec, M. A.; Coggins, B. E.; Isern, N. G. 3D TOCSY-HSQC NMR for metabolic flux analysis using non-uniform sampling. *Anal. Chem.* **2016**, *88*, 2825−2831.

DOI: 10.1021/acs.jproteome.7b00457
J. Proteome Res. 2017, 16, 3774−3786

(29) Kazimierczuk, K.; Orekhov, V. Non-uniform sampling: post-Fourier era of NMR data collection and processing. *Magn. Reson. Chem.* **2015**, *53*, 921−926.
(30) Li, D. W.; Wang, C.; Brüschweiler, R. Maximal clique method for the automated analysis of NMR TOCSY spectra of complex mixtures. *J. Biomol. NMR* **2017**, *68*, 195−202.
(31) Pence, H. E.; Williams, A. ChemSpider: an online chemical information resource. *J. Chem. Educ.* **2010**, *87*, 1123−1124.
(32) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. METLIN: a metabolite mass spectral database. *Ther. Drug Monit.* **2005**, *27*, 747−751.
(33) Abraham, R. J.; Reid, M. Proton chemical shifts in NMR. Part 16. proton chemical shifts in acetylenes and the anisotropic and steric effects of the acetylene group. *J. Chem. Soc., Perkin Trans. 2* **2001**, 1195−1204.
(34) Munkres, J. Algorithms for the assignment and transportation problems. *J. Soc. Ind. Appl. Math.* **1957**, *5*, 32−38.
(35) Stahl, M.; Mauser, H. Database clustering with a combination of fingerprint and maximum common substructure methods. *J. Chem. Inf. Model.* **2005**, *45*, 542−548.
(36) Raymond, J. W.; Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 521−533.
(37) Sumner, L. W.; Amberg, A.; Barrett, D.; Beale, M. H.; Beger, R.; Daykin, C. A.; Fan, T. W.; Fiehn, O.; Goodacre, R.; Griffin, J. L.; Hankemeier, T.; Hardy, N.; Harnly, J.; Higashi, R.; Kopka, J.; Lane, A. N.; Lindon, J. C.; Marriott, P.; Nicholls, A. W.; Reily, M. D.; Thaden, J. J.; Viant, M. R. Proposed minimum reporting standards for chemical analysis. *Metabolomics* **2007**, *3*, 211−221.
(38) Hoffmann, F.; Li, D. W.; Sebastiani, D.; Brüschweiler, R. Improved quantum chemical NMR chemical shift prediction of metabolites in aqueous slution toward the validation of unknowns. *J. Phys. Chem. A* **2017**, *121*, 3071−3078.
(39) Barry, G. T.; Roark, E. L-Fucosamine and 4-oxo-norleucine as constituents in mucopolysaccharides of certain enteric bacteria. *Nature* **1964**, *202*, 493−494.
(40) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. NMRPipe - a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **1995**, *6*, 277−293.
(41) Palcic, B.; Garner, D.; Susnik, B. UCSF Sparky-an NMR display, annotation and assignment tool. *J. Cell. Biochem.* **1993**, *53*, 17c, 254.
(42) Guan, S. H.; Marshall, A. G. Stored waveform inverse Fourier transform (SWIFT) ion excitation in trapped-ion mass spectometry: Theory and applications. *Int. J. Mass Spectrom. Ion Processes* **1996**, *157*, 5−37.
(43) Ledford, E. B.; Rempel, D. L.; Gross, M. L. Space charge effects in Fourier transform mass spectrometry. II. Mass Calibration. *Anal. Chem.* **1984**, *56*, 2744−2748.
(44) Shi, S. D. H.; Drader, J. J.; Freitas, M. A.; Hendrickson, C. L.; Marshall, A. G. Comparison and interconversion of the two most common frequency-to-mass calibration functions for Fourier transform ion cyclotron resonance mass spectrometry. *Int. J. Mass Spectrom.* **2000**, *195*, 591−598.

DOI: 10.1021/acs.jproteome.7b00457
J. Proteome Res. 2017, 16, 3774−3786