Published in final edited form as:

Anal Chem. 2015 April 7; 87(7): 3864–3870. doi:10.1021/ac504633z.

Metabolomics Beyond Spectroscopic Databases: A Combined MS/NMR Strategy for the Rapid Identification of New Metabolites in Complex Mixtures

Kerem Bingol¹, Lei Bruschweiler-Li², Cao Yu², Arpad Somogyi², Fengli Zhang³, and Rafael Bruschweiler^{1,2,3,*}

¹Department of Chemistry and Biochemistry, The Ohio State University, Columbus, Ohio 43210, United States

²Campus Chemical Instrument Center, The Ohio State University, Columbus, Ohio 43210, United States

³National High Magnetic Field Laboratory, Florida State University, Tallahassee, Florida 32310, United States

Abstract

A novel strategy is introduced that combines high-resolution mass spectrometry (MS) with NMR for the identification of unknown components in complex metabolite mixtures encountered in metabolomics. The approach first identifies the chemical formulas of the mixture components from accurate masses by MS and then generates all feasible structures (structural manifold) that are consistent with these chemical formulas. Next, NMR spectra of each member of the structural manifold are predicted and compared with the experimental NMR spectra in order to identify the molecular structures that match the information obtained from both the MS and NMR techniques. This combined MS/NMR approach was applied to *E. coli* extract where the approach correctly identified a wide range of different types of metabolites, including amino acids, nucleic acids, polyamines, nucleosides and carbohydrate conjugates. This makes this approach, which is termed SUMMIT MS/NMR, well suited for high-throughput applications for the discovery of new metabolites in biological and biomedical mixtures overcoming the need of experimental MS and NMR metabolite databases.

Graphical Abstract

*To whom correspondence should be addressed: Rafael Bruschweiler, Ph.D., Newman & Wolfrom Laboratory of Chemistry, Department of Chemistry and Biochemistry, The Ohio State University, Columbus, Ohio 43210, bruschweiler.1@osu.edu.

Supporting Information

Figures of mass spectra and 2D HSQC NMR spectrum of a model mixture and *E. coli* cell lysate along with tables of chemical formulas and NMR peak lists of mixture compounds. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

Introduction

Metabolomics as a field of research has gained significant attention over the recent past as it is developing rapidly into a powerful way to comprehensively study complex biological systems from a small molecule perspective.¹² Small biological molecules or metabolites are the key players of metabolism, which makes the analysis of their chemical structure and their abundance important as they are direct indicators of the phenotype of the state of a biological system, such as an organism, organ, or biofluid.³⁴

Mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy are the two most powerful experimental methods for metabolomics. This is because of the excellent resolution power that both of these techniques can provide to detect individual molecular species.⁵⁻⁶ Unfortunately, detection alone does not always lead to the unambiguous identification of metabolites.⁷ In fact, many of the signals found in NMR and MS spectra of metabolic samples belong to molecules whose identification is notably hard. Identification of these unknown molecules has been recognized as a central bottleneck hampering progress in the field of metabolomics.⁸

Despite the individual power of the MS and NMR methods, the synergistic use of these two methods has turned out to be remarkably challenging, which in part is because their information content is too complementary to be combined in a straightforward manner. Methods have been introduced that integrate NMR and MS by means of multivariate statistical analysis applied to a large number of samples.⁹⁻¹¹ Such approaches correlate NMR signals with masses, but they do not provide molecular structures.

Metabolite identification by NMR is usually performed in two steps. In a first step, the NMR spectrum of the metabolite mixture is deconvoluted into single resonances or groups of resonances that belong to an individual component.¹² In the second step, these spectral fingerprints are queried against one or several NMR metabolite databases. The success of this approach for the positive identification of a metabolite depends not only on the quality of the spectral deconvolution, but it also requires the presence of the NMR spectrum of the compound in the database, measured under similar or identical conditions as the mixture. Although excellent progress has been made in the compilation of NMR metabolomics databases, such as MMCD,¹³ BMRB,¹⁴ HMDB,¹⁵ and COLMAR,¹⁶⁻¹⁸ the further expansion of these databases is time and labor-intensive. The current databases typically contain hundreds of metabolites, whereas the number of different metabolites in a single organism has been estimated to be in the thousands.¹⁹ Therefore, approaches that rely on

databases have clear limitations when it comes to the determination of the entire metabolome of a complex biological system. Recently, 2D NMR spectroscopy has been used for the characterization of the backbone topologies of unknown molecules in metabolomics samples toward the elucidation of metabolite structures in complex mixtures.^{20,21} In this way, it was possible to extract 112 carbon backbone topologies from a single *E. coli* cell lysate.²¹

Identification of metabolites by MS faces other challenges. Detection of the accurate mass of a compound permits the determination of its chemical formula, but the number of molecular structures with the same formula grows exponentially with the molecular weight.²² To address this degeneracy, additional information is required such as the one obtained from MS/MS fragments,²³ where the fragment masses are used as fingerprints for the identification of the specific structures by comparing them with fragmentation patterns of known compounds stored in databases, such as METLIN²⁴ and HMDB.¹⁵ This approach is of limited use for the *de novo* identification of mixture compounds, since only compounds can be identified whose fragmentation patterns are already contained in such databases.

Traditionally, identification of unknown, i.e. uncatalogued, metabolites requires their isolation through time-consuming purification from complex mixtures by using separation techniques, such as chromatography, followed by extensive characterization by NMR, MS, X-ray, and other techniques.^{25,26} The utility of this approach is limited in the context of high-throughput applications and, in addition, the purification steps may result in a significant decrease in metabolite concentration rendering *de novo* structure elucidation a challenge because of insufficient sensitivity.

Here, we propose a metabolite identification strategy of complex mixtures by combining MS with NMR in a novel way. It neither requires purification nor the use of NMR and MS metabolite databases. This makes the method suitable for high-throughput identification of new metabolites. We term this approach SUMMIT MS/NMR for “Structure of Unknown Metabolomic Mixture components by MS/NMR”.

Experimental Section

Sample preparation

A model mixture was prepared in 50%/50% (v/v) ACN/H₂O with 0.1% formic acid by adding 10 metabolites: lysine, shikimate, carnitine, isoleucine, glutamate, histidine, arginine, alanine, ornithine, and glutamine. The final concentration of each metabolite was 10 μ M.

E. coli BL21(DE3) cells were cultured at 37 °C, at 250 rpm in M9 minimum medium with glucose (natural abundance, 5g/L) added as sole carbon source. One liter of culture at OD ~3 was centrifuged at 5000xg for 20 min at 4 °C, and the cell pellet was resuspended in 50 mL of 50 mM phosphate buffer at pH 7.0. Cell suspension was then subjected to centrifugation for cell pellet collection. The cell pellet was resuspended in 10 mL of ice cold water and exposed to freeze-thaw procedure 3 times. The sample was centrifuged at 20000xg at 4°C for 15 min to remove the cell debris. Pre-chilled methanol and chloroform were sequentially added to the supernatant under vigorous vortex at H₂O:methanol:chloroform ratios of 1:1:1

(v/v/v). The mixture was then left at -20°C overnight for phase separation. Next, it was centrifuged at $4000\times g$ for 20 min at 4°C , and the clear top hydrophilic phase was collected and subjected to rotary evaporator processing to have the methanol content reduced. Finally, the sample is lyophilized. The dry sample is then divided into two parts, one for MS and one for NMR analysis. The NMR sample is prepared by dissolving the material in $\sim 200\ \mu\text{L}$ D_2O , which is then transferred to a 3-mm NMR tube. The MS sample is dissolved $200\ \mu\text{L}$ H_2O , $10\ \mu\text{L}$ of that is diluted 10 folds by 50%/50% (v/v) ACN/ H_2O with 0.1% formic acid. The resulting solution is centrifuged at 13000 rpm 4°C for 5 min and the supernatant is used for direct infusion MS.

NMR experiments and processing

The 2D ^{13}C - ^1H HSQC²⁷ spectra of the ten-compound model mixture were downloaded from the BMRB database. All NMR spectra of *E. coli* cell lysate were collected using a Bruker AVANCE solution-state NMR spectrometer equipped with a cryogenically cooled probe at 800 MHz proton frequency at 298 K. 2D ^{13}C - ^1H HSQC²⁷ and 2D ^{13}C - ^1H HSQC-TOCSY²⁸ spectra of *E. coli* cell lysate were collected with $N_1=512$ and $N_2=1024$ complex points. The spectral width along the indirect and the direct dimensions were 34209.9 Hz and 8802.8 Hz, respectively. The number of scans per t_1 increment was 64. The transmitter frequency offset were 85 ppm in the ^{13}C dimension and 4.7 ppm in the ^1H dimension. TOCSY mixing time for 2D ^{13}C - ^1H HSQC-TOCSY was set to 90 ms. The total measurement time for each experiment was 36 hours. 2D ^1H - ^1H TOCSY²⁹ spectrum of *E. coli* cell lysate was collected with $N_1=512$ and $N_2=1024$ complex points. The spectral widths along the indirect and the direct dimensions were both 8802.8 Hz. The number of scans per t_1 increment was 8. The transmitter frequency offset were 4.7 ppm in both ^1H dimensions. TOCSY mixing time was set to 90 ms. The total measurement time was 12 hours. 2D ^{13}C - ^1H HMBC³⁰ spectrum of *E. coli* cell lysate was collected with $N_1=768$ and $N_2=2048$ complex points. The spectral width along the indirect and the direct dimensions were 50310.8 Hz and 8802.8 Hz, respectively. The number of scans per t_1 increment was 32. The transmitter frequency offset was 125 ppm in the ^{13}C dimension and 4.7 ppm in the ^1H dimension. The total measurement time was about 30 hours. Data were zero-filled, Fourier transformed, and phase and baseline corrected using NMRPipe.³¹

Mass spectrometry experiments and processing

Direct infusion studies were conducted in positive ion mode detection on a Bruker maXis 4G ESI Q-TOF instrument (electrospray ionization quadrupole time-of-flight mass spectrometer). The instrument was calibrated with Agilent Low-Concentration Tuning Mix (Part No. G1969-85000) before sample analysis achieving a mass accuracy of ± 5 ppm. The samples were directly infused to the ESI source at $2\ \mu\text{L}/\text{min}$. The settings for the Q-TOF mass spectrometer were as follows: capillary voltage, 4500 V; end plate offset, -500 V; drying gas flow (N_2), $4.0\ \text{L}/\text{min}$; drying gas temperature, 200°C ; and nebulizer gas (N_2), 0.5 bar.

Predicted NMR database generation and HSQC comparison

Molecular formulas were searched in ChemSpider database³² by tolerating plus or minus one hydrogen mass in each formula: for example, $\text{C}_7\text{H}_{15}\text{NO}_3$ is searched as $\text{C}_7\text{H}_{14-16}\text{NO}_3$.

The following six elements were considered for the generation of the molecular formula from exact masses: C, H, N, O, P, and S. 2D ^{13}C - ^1H HSQC spectra of the returned structures are predicted by using MestReNova 9.0.1 (Mestrelab Research, Santiago de Compostela, Spain). HSQC prediction of each molecule takes about 10 seconds on a desktop computer. The comparison of each HSQC peak list of the experimental NMR spectra is performed by using the query algorithm of COLMAR ^{13}C - ^1H HSQC web server.³³

Results and Discussion

New MS/NMR strategy

The general workflow of SUMMIT MS/NMR strategy for the identification of metabolites is illustrated in Figure 1. For a sample of a metabolite mixture of unknown composition the high-resolution mass spectrum is determined using, e.g., a Q-TOF, Orbitrap, or FTICR MS. Accurate masses of the components are extracted from the mass spectrum and converted to unique molecular formulas. For each molecular formula, all possible structures are generated, which we call the ‘structural manifold’ of all molecular formulas of a mixture. The structural manifold can be very large and include hundreds to thousands of different structures. The NMR spectrum (chemical shifts) of each structure is then predicted and stored. Meanwhile, the experimental NMR spectrum is determined for the same mixture and deconvoluted into the NMR spectra of individual components. The NMR spectrum of each component is compared to the predicted NMR spectra of the total structural manifold, which is the combination of all structural manifolds, and the structures are rank-ordered according to the level of agreement. In this way, the NMR spectrum together with the predicted chemical shifts are used as a “filter” to identify those molecular structures that are most consistent with all available NMR and MS data.

Application to ten-compound model mixture

First the method was tested on a ten-compound metabolite mixture consisting of carnitine, arginine, isoleucine, ornithine, lysine, glutamate, glutamine, alanine, histidine, and shikimate, which was analyzed by Q-TOF MS. From the resulting direct injection mass spectrum (see Supporting Information Figure S1), we picked the 50 largest peaks by height, which resulted in 22 unique molecular formulas shown in Supplementary Table S1. For each molecular formula all feasible structures were determined with ChemSpider.³² For example, for the molecular formula $\text{C}_7\text{H}_{15}\text{NO}_3$, ChemSpider returned a structural manifold comprising 362 different chemical structures one of which is carnitine. According to ChemSpider, for the 22 molecular formulas there exist a total of 4772 different structures constituting the total structural manifold. 2D NMR ^{13}C - ^1H HSQC²⁷ spectra of the structures were predicted by using the MestReNova software. Experimental ^{13}C - ^1H HSQC spectra of the 10 compounds (Table S2) were compared one-by-one against the predicted HSQC spectra of the 4772 structures using a scoring function that is analogous to the one used for the querying of HSQC spectra against our COLMAR ^{13}C - ^1H HSQC web server.³³ Lysine, carnitine, histidine, alanine, ornithine, and glutamine were returned as the top hits among the 4772 structures. Isoleucine, arginine, and glutamate were returned as the second best hits among the 4772 structures. For these 3 molecules, the top hits are structurally very similar to the true structures (Figure S2). The 10th metabolite, the acidic shikimate, was not detected in

the mass spectrum, presumably because the mass spectrometer was operated in positive ion mode. A summary of the results is given in Table 1. Overall, the application of SUMMIT MS/NMR to the model mixture shows the potential of this approach to determine molecular structures present in complex mixtures without the use of NMR and MS metabolomics databases.

Application to *E. coli* polar cell extract

The approach was then applied to an actual metabolomics sample, namely an *E. coli* cell extract, which was injected into the Q-TOF mass spectrometer. From the resulting MS spectrum 56 unique molecular formulas could be extracted where the majority of them belong to the 500 highest MS signals (Fig. S3 and Table S3). For each molecular formula all feasible structures were determined using ChemSpider, resulting in a total structural manifold of 13,872 structures. 2D ^{13}C - ^1H HSQC spectrum of each of these structures was predicted by using MestReNova software. Meanwhile, high-resolution 2D ^{13}C - ^1H HSQC,²⁷ 2D ^1H - ^1H TOCSY,²⁹ 2D ^{13}C - ^1H HSQC-TOCSY²⁸ and 2D ^{13}C - ^1H HMBC³⁰ spectra of the *E. coli* sample were acquired from the same *E. coli* sample material. The 2D ^{13}C - ^1H HSQC spectrum was then deconvoluted into subspectra belonging to individual mixture components by using connectivity information derived from 2D ^1H - ^1H TOCSY, 2D ^{13}C - ^1H HSQC-TOCSY, and 2D ^{13}C - ^1H HMBC spectra (Fig. S4). The chemical shift list (cross-peak list) of the deconvoluted ^{13}C - ^1H HSQC subspectra of the *E. coli* extract (Table S4) was quantitatively compared one-by-one against the peak lists predicted for each of the 13,872 manifold structures using a scoring function that is analogous to the one used for the querying of HSQC spectra against COLMAR ^{13}C - ^1H HSQC web server. This procedure is exemplified for N-acetylputrescine, aspartate, and nicotinate in Figure 2. Aspartate, alanine, betaine, GABA, glutamine, arginine, lysine, methionine, N-acetylputrescine, spermidine, tyrosine, threonine, uracil, and nicotinate were returned as the top hits among the 13,872 structures. Isoleucine and phenylalanine were returned as second best hits, while adenosine, glutamate, leucine, and valine were returned as third-best hits among all 13,872 structures. A summary of the results is given in Table 2. In most cases, false positive structures, which were returned as best hits, are structurally very similar to the true structures (Fig. S5). These results clearly demonstrate the power of NMR chemical shift information as an effective filter to identify the correct structures among the large structural manifold belonging to MS-derived molecular formulas.

The novel metabolite identification strategy introduced here enables accurate high-throughput applications for the identification of unknown metabolites because of two main reasons. First, the approach does not rely on experimental NMR or MS databases. Its ability to identify molecules is therefore not bound by the limited number of metabolites contained in current databases. Secondly, it can provide direct identification of crude metabolic extracts, with little or no purification, as shown here for *E. coli*. Other, conceptually related approaches have been introduced previously for structure elucidation of pure compounds by combining MS and NMR.³⁴ However, in order to study a complex mixture in this way, one would need to purify each component of interest (e.g. by HPLC), then collect MS and NMR spectra of each pure compound and apply these methods toward structure elucidation. This makes them impractical for the analysis of a potentially large number of compounds present

in complex mixtures, such as those encountered in metabolomics. SUMMIT MS/NMR has been designed to overcome this challenge to be able to perform structure elucidation directly in complex mixtures as is demonstrated in the following.

The goal of this study is to provide a proof-of-principle of SUMMIT MS/NMR and, therefore, it was applied to the identification of metabolites that were already known by other means. The method proved instrumental during identification of metabolites in *E. coli* extract. For example, when we started analyzing the *E. coli* metabolome by NMR, N-acetylputrescine signals could not be assigned, because the NMR spectrum of this compound was not present in the BMRB and COLMAR databases used. Although the NMR spectrum of N-acetylputrescine was present in the HMDB, upon querying the HMDB with our input data, it did not return N-acetylputrescine. By contrast SUMMIT MS/NMR positively identified N-acetylputrescine without spectroscopic database information and its presence in *E. coli* was verified after visual comparison with the HMDB entry of this molecule. This illustrates the potential of the SUMMIT MS/NMR. The identification of truly unknown metabolites by SUMMIT MS/NMR in various systems is currently under way in our lab.

A main requirement for the success of SUMMIT MS/NMR is that a metabolite of interest is detected both by MS and NMR. Therefore, a metabolite should be present at least at low micromolar concentration to be detected in NMR experiments. Furthermore, as is the case in any MS application, ionization of metabolites in the mass spectrometer is critically important. We found a substantial number of metabolites that are well detectable by both NMR and MS. These are the metabolites that can be targeted by the SUMMIT MS/NMR approach. If a metabolite does not ionize, such as shikimate of the model mixture in positive mode, the NMR chemical shifts of shikimate did not lead to a false positive hit, because the chemical shift differences of shikimate with respect to the MS-derived manifold were much larger than the differences of the true positive hits. We attribute this to the fact that most metabolites have multiple NMR resonances, which drastically reduces the likelihood of an accidentally low chemical shift difference.

In the current study, we used standard direct infusion electrospray ionization (ESI) mass spectrometry for a proof of concept of the method, but more sophisticated MS approaches can be applied to further optimize the number of metabolites detected by mass spectrometry. This includes combined chromatographic techniques, such as LC-MS,³⁵ provided that the MS part of the LC-MS instrument is a high-resolution mass spectrometer such as a Q-TOF, Orbitrap, or FTICR MS. On the NMR side, there are a number of techniques now available for the increase of resolution or the reduction of NMR time, such as NUS sampling,³⁶ which is most useful when sampling and not sensitivity is the limiting factor.

For a proof-of-principle demonstration, we only considered $[M+H]^+$ ions in positive ion mode. However, metabolites can show up as adducts ($[M+Na]^+$, $[M+K]^+$, etc.) and fragments (e.g., $[M+H-H_2O]^+$) in positive ion mode mass spectra or, similarly, some metabolites can show up in the negative mode. Therefore application of SUMMIT MS/NMR to molecular formulas corresponding to these m/z values will be a natural extension of this method. However, grouping of different adduct and fragment features of the same molecule

before applying the SUMMIT MS/NMR approach, e.g. using recently developed software,³⁷ should further increase the efficiency and accuracy of the approach. We used a Q-TOF instrument for the extraction of unique formulas from accurate masses with less than 5 ppm (parts per million) m/z determination error. Determination of a molecular formula for each detected MS signal was not possible, because the MS signals either did not correspond to any molecular formula or they corresponded to more than one molecular formula (within experimental error). Although for the latter cases our NMR filter could still identify true structures, isotopic abundance patterns in mass spectrometry and/or higher resolution mass spectrometers, such as FTICR MS, can be analyzed to extract unique formulas from detected masses.³⁸ The availability of a set of high-quality molecular formulas at the beginning of the SUMMIT MS/NMR procedure is important to ensure the accuracy of this approach.

For the majority of cases encountered in this work, comparison of ^{13}C - ^1H HSQC signals provided a sufficient amount of complementary information to extract the true structures from the structural manifolds. For those cases where the true structure was not the top hit, additional NMR-derived information can be used. For example, leucine was not identified as the top hit (Fig. S5); however, in this case the false positive top hit can easily be excluded by comparing the NMR TOCSY pattern expected for this compound with the experimental TOCSY pattern measured for leucine. For this additional step, neither physical isolation of the metabolite of interest nor prior chemical knowledge about the sample composition or the unknown metabolite are required, because the experimental TOCSY pattern of an unknown molecule is available from a 2D ^1H - ^1H TOCSY or a 2D ^{13}C - ^1H HSQC-TOCSY spectrum of the same complex mixture (i.e., *E. coli* cell lysate). Molecular topology²¹, NMR intensities, MS/MS pattern and retention time are additional constraints that can be used to identify false positives.

The structural manifolds were constructed with ChemSpider.³² Since ChemSpider does not contain all theoretically possible structures, one can further increase the structural manifold by combining multiple databases such as CAS and PubChem or one can use computer-generated structures by software such as MOLGEN.²² 1D ^{13}C NMR spectroscopy has already been used to optimize structures predicted by such software.³⁹

The SUMMIT MS/NMR approach requires the prediction of all chemical shifts of the structural manifold, which can be considered a “database on the fly”. However, the approach does not rely on repositories of experimental NMR or MS data of previously identified and characterized metabolites. Therefore, SUMMIT MS/NMR can be used to identify new compounds that are not contained in existing databases.

Conclusion

SUMMIT MS/NMR is a novel approach for metabolite identification that combines the highly complementary information provided by NMR and mass spectrometry. Most metabolomics applications deal with samples that contain many metabolites with different molecular formulas. Since it is not known in advance which NMR signals belong to which molecular formula in the complex mixture, traditionally a purification step for each metabolite is required, which limits high-throughput applications. In this study experimental

NMR signals are compared with the NMR spectra that are predicted from the structures of all detected molecular formulas. In this way, NMR chemical shifts are used as a potent ‘orthogonal’ filter to extract correct molecular structures from large manifolds of accurate molecular masses and the corresponding molecular formulas without any purification of each metabolite from complex mixture. This opens up new avenues for high-throughput identification of potentially unknown metabolites overcoming fundamental limitations of database approaches in the search of new metabolites in complex biological mixtures. Finally, it is worth to emphasize that SUMMIT MS/NMR approach is applicable to mixtures of a broad range of origins ranging from biological to biomedical systems, synthetic mixtures, and nutrition.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institutes of Health (grant R01 GM 066041).

References

1. Nicholson JK, Holmes E, Kinross JM, Darzi AW, Takats Z, Lindon JC. *Nature*. 2012; 491:384–392. [PubMed: 23151581]
2. Bingol K, Brüschweiler R. *Anal Chem*. 2014; 86:47–57. [PubMed: 24195689]
3. Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L. *Nat Biotechnol*. 2000; 18:1157–1161. [PubMed: 11062433]
4. Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ, Westerhoff HV, van Dam K, Oliver SG. *Nat Biotechnol*. 2001; 19:45–50. [PubMed: 11135551]
5. Lenz EM, Wilson ID. *J Proteome Res*. 2007; 6:443–458. [PubMed: 17269702]
6. Dettmer K, Aronov PA, Hammock BD. *Mass Spectrom Rev*. 2007; 26:51–78. [PubMed: 16921475]
7. Fiehn O. *Plant Mol Biol*. 2002; 48:155–171. [PubMed: 11860207]
8. Wishart DS. *Bioanalysis*. 2011; 3:1769–1782. [PubMed: 21827274]
9. Crockford DJ, Holmes E, Lindon JC, Plumb RS, Zirah S, Bruce SJ, Rainville P, Stumpf CL, Nicholson JK. *Anal Chem*. 2006; 78:363–371. [PubMed: 16408915]
10. Pan ZZ, Gu HW, Talaty N, Chen HW, Shanaiah N, Hainline BE, Cooks RG, Raftery D. *Anal Bioanal Chem*. 2007; 387:539–549. [PubMed: 16821030]
11. Marshall D, Lei S, Worley B, Huang Y, Garcia-Garcia A, Franco R, Dodds E, Powers R. *Metabolomics*. 2014:1–12.
12. Bingol K, Brüschweiler R. *Anal Chem*. 2011; 83:7412–7417. [PubMed: 21848333]
13. Cui Q, Lewis IA, Hegeman AD, Anderson ME, Li J, Schulte CF, Westler WM, Eghbalnia HR, Sussman MR, Markley JL. *Nat Biotechnol*. 2008; 26:162–164. [PubMed: 18259166]
14. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Wenger RK, Yao HY, Markley JL. *Nucl Acids Res*. 2008; 36:D402–D408. [PubMed: 17984079]
15. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S, Mandal R, Sinelnikov I, Xia JG, Jia L, Cruz JA, Lim E, Sobsey CA, Shrivastava S, Huang P, Liu P, Fang L, Peng J, Fradette R, Cheng D, Tzur D, Clements M, Lewis A, De Souza A, Zuniga A, Dawe M, Xiong YP, Clive D, Greiner R, Nazzyrova A, Shaykhtudinov R, Li L, Vogel HJ, Forsythe I. *Nucl Acids Res*. 2009; 37:D603–D610. [PubMed: 18953024]

16. Robinette SL, Zhang FL, Bruschweiler-Li L, Brüschweiler R. *Anal Chem.* 2008; 80:3606–3611. [PubMed: 18422338]
17. Bingol K, Zhang F, Bruschweiler-Li L, Brüschweiler R. *Anal Chem.* 2012; 84:9395–9401. [PubMed: 23016498]
18. Bingol K, Bruschweiler-Li L, Li DW, Brüschweiler R. *Anal Chem.* 2014; 86:5494–5501. [PubMed: 24773139]
19. Guo AC, Jewison T, Wilson M, Liu YF, Knox C, Djoumbou Y, Lo P, Mandal R, Krishnamurthy R, Wishart DS. *Nucl Acids Res.* 2013; 41:D625–D630. [PubMed: 23109553]
20. Zhang FL, Bruschweiler-Li L, Brüschweiler R. *J Am Chem Soc.* 2010; 132:16922–16927. [PubMed: 21062057]
21. Bingol K, Zhang F, Bruschweiler-Li L, Brüschweiler R. *J Am Chem Soc.* 2012; 134:9006–9011. [PubMed: 22540339]
22. Benecke C, Grund R, Hohberger R, Kerber A, Laue R, Wieland T. *Analytica Chimica Acta.* 1995; 314:141–147.
23. Tautenhahn R, Cho K, Uritboonthai W, Zhu ZJ, Patti GJ, Siuzdak G. *Nat Biotechnol.* 2012; 30:826–828. [PubMed: 22965049]
24. Zhu ZJ, Schultz AW, Wang JH, Johnson CH, Yannone SM, Patti GJ, Siuzdak G. *Nat Protocols.* 2013; 8:451–460. [PubMed: 23391889]
25. Koehn FE, Carter GT. *Nature Rev Drug Discov.* 2005; 4:206–220. [PubMed: 15729362]
26. Corcoran O, Spraul M. *Drug Discov Today.* 2003; 8:624–631. [PubMed: 12867148]
27. Bodenhausen G, Ruben DJ. *Chem Phys Lett.* 1980; 69:185–189.
28. Lerner L, Bax A. *J Magn Reson.* 1986; 69:375–380.
29. Braunschweiler L, Ernst RR. *J Magn Reson.* 1983; 53:521–528.
30. Bax A, Summers MF. *J Am Chem Soc.* 1986; 108:2093–2094.
31. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. *J Biomol NMR.* 1995; 6:277–293. [PubMed: 8520220]
32. Pence HE, Williams A. *J Chem Educ.* 2010; 87:1123–1124.
33. Bingol K, Li DW, Bruschweiler-Li L, Cabrera OA, Megraw T, Zhang F, Brüschweiler R. *ACS Chem Biol.* 2014; doi: 10.1021/cb5006382
34. Elyashberg M, Blinov K, Molodtsov S, Smurnyy Y, Williams AJ, Churanova T. *J Chem Inf.* 2009; 1:3.
35. Lin L, Yu QA, Yan XM, Hang W, Zheng JX, Xing JC, Huang BL. *Analyst.* 2010; 135:2970–2978. [PubMed: 20856980]
36. Billeter, M.; Orekhov, VY. *Novel Sampling Approaches in Higher Dimensional NMR.* Springer; Heidelberg, Germany: 2012.
37. Kuhl C, Tautenhahn R, Bottcher C, Larson TR, Neumann S. *Anal Chem.* 2012; 84:283–289. [PubMed: 22111785]
38. Kind T, Fiehn O. *BMC Bioinformatics.* 2006; 7
39. Meiler J, Will M. *J Chem Inf Comp Sci.* 2001; 41:1535–1546.

Figure 1.

Schematic representation of the SUMMIT MS/NMR strategy for the identification of metabolites in complex metabolomic mixtures by the combined use of mass spectrometry and 1D NMR spectroscopy. High-resolution MS yields the unique molecular formulas of the metabolites present in the mixture (left). For each molecular formula, all possible structures are generated representing the total 'structural manifold' depicted as the sum of the three local manifolds (green, red, blue; middle) each belonging to a different mass. Next, NMR chemical shifts are predicted for all manifold structures. Comparison of the predicted with the experimental NMR chemical shifts (right) allows identification of the structures that are present in the mixture, requiring neither an NMR nor an MS metabolomics database.

Figure 2.

Application of the SUMMIT MS/NMR method to an *E. coli* cell lysate with 2D NMR. High-resolution MS yields the unique molecular formulas of the metabolites present in the lysate. From the total structural manifold belonging to these masses, the 2D ^{13}C - ^1H HSQC spectrum is predicted for each structure. Meanwhile, an experimental 2D ^{13}C - ^1H HSQC spectrum of the lysate is deconvoluted into ^{13}C - ^1H HSQC chemical shifts of each metabolite by combining information from 2D NMR experiments. Comparison of the experimental ^{13}C - ^1H HSQC chemical shifts of each metabolite with the predicted ^{13}C - ^1H HSQC spectra for each of the manifold structures allows the unique identification of the

metabolites belonging to detected molecular formulas as is illustrated for N-acetylputrescine, aspartate, and nicotinate.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

SUMMIT MS/NMR results for ten-compound model mixture

Metabolite	Rank ^a	¹ H ^b	¹³ C ^c	m/z ^d	Size ^e
lysine	1	0.099	1.338	147.1124	294
carnitine	1	0.068	2.760	162.1123	362
histidine	1	0.221	1.823	156.0763	770
alanine	1	0.141	1.004	90.0552	74
ornithine	1	0.135	1.143	133.0972	241
glutamine	1	0.085	1.447	147.0760	176
isoleucine	2	0.093	2.283	132.1017	535
arginine	2	0.163	1.697	175.1187	65
glutamate	2	0.142	3.164	148.0600	167

^a Rank ordered agreement between experimental and predicted HSQC spectra of a given metabolite. For example, after comparison of the experimental HSQC spectrum of lysine with the predicted HSQC spectrum of each of the 4772 structures constituting the total structural manifold, it is found that the predicted HSQC spectrum of lysine itself is most similar to the experimental HSQC spectrum and therefore has rank 1.

^b Average ¹H chemical shift difference (in units of ppm) between the experimental and predicted chemical shifts.

^c Average ¹³C chemical shift difference (in units of ppm) between the experimental and the predicted chemical shifts.

^d [M+H]⁺ (= monoisotopic mass + proton mass) m/z detected in the mass spectrum.

^e Number of structures for a given chemical formula (obtained with ChemSpider).

Table 2

SUMMIT MS/NMR results for *E. coli* cell lysate

Metabolite	Rank ^a	¹ H ^b	¹³ C ^c	m/z ^d	Size ^e
aspartate	1	0.104	2.123	134.0447	56
alanine	1	0.154	0.958	90.0553	75
betaine	1	0.121	1.749	118.0864	333
GABA	1	0.058	2.949	104.0712	164
glutamine	1	0.089	1.492	147.0765	176
arginine	1	0.163	1.726	175.1187	65
lysine	1	0.098	1.405	147.1127	295
methionine	1	0.127	1.241	150.0588	136
N-acetylputrescine	1	0.164	1.948	131.1180	383
spermidine	1	0.191	1.387	146.1652	31
tyrosine	1	0.105	1.805	182.0809	999
threonine	1	0.088	2.028	120.0650	142
uracil	1	0.112	0.291	113.0349	97
nicotinate	1	0.117	3.003	124.0393	89
isoleucine	2	0.093	2.325	132.1020	535
phenylalanine	2	0.122	2.884	166.0861	1161
adenosine	3	0.189	2.467	268.1041	340
glutamate	3	0.138	3.122	148.0605	167
leucine	3	0.203	3.899	132.1020	535
valine	3	0.089	0.852	118.0864	333
putrescine	6	0.231	1.848	89.1077	40

^a Rank ordered agreement between experimental and predicted HSQC spectra of a given metabolite.

^b Average ¹H chemical shift difference (in units of ppm) between the experimental and predicted chemical shifts.

^c Average ¹³C chemical shift difference (in units of ppm) between the experimental and the predicted chemical shifts.

^d [M+H]⁺ (= monoisotopic mass + proton mass) m/z detected in the mass spectrum.

Number of structures for a given chemical formula (obtained with ChemSpider).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript