NMR AS A TOOL FOR COMPOUND IDENTIFICATION IN MIXTURES

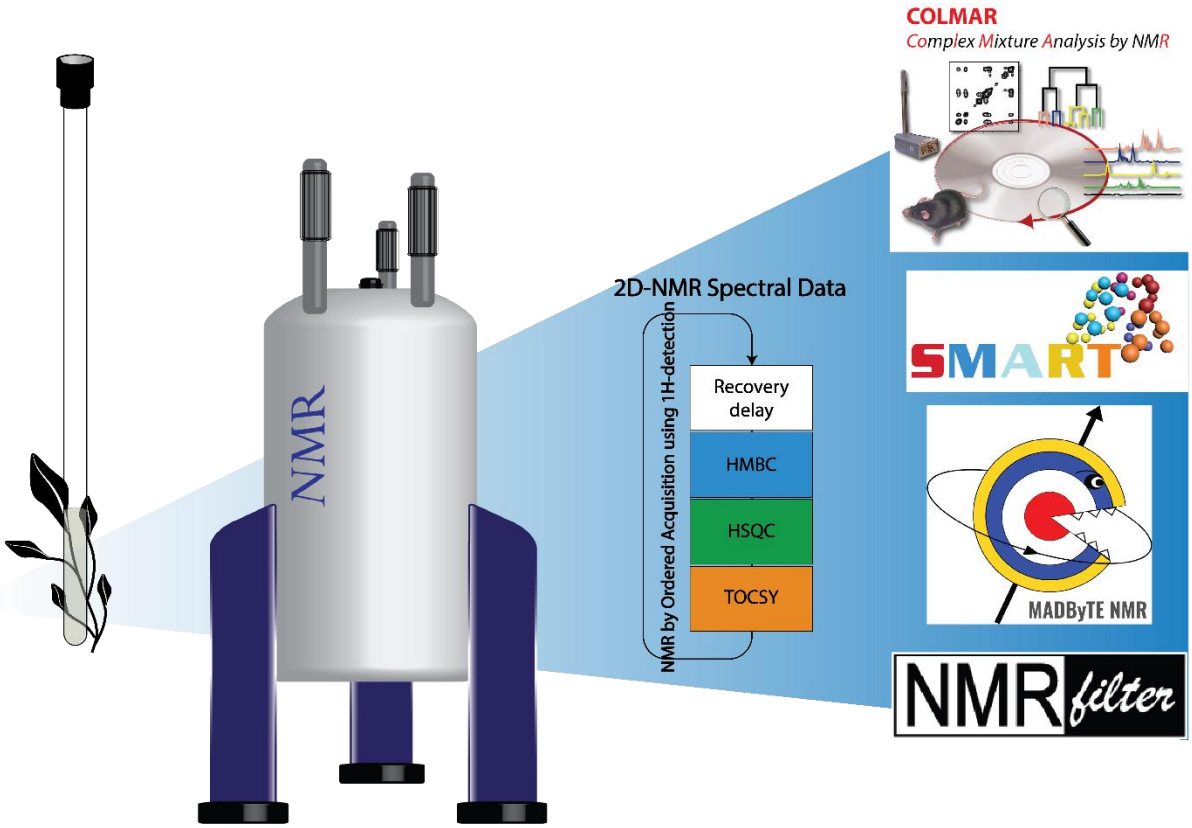
Ricardo M. Borges*[1], Gabriela de Assis Ferreira[1], Mariana Martins Campos[1], Andrew Magno Teixeira[1], Fernanda das Neves Costa[1], Fernanda Oliveira das Chagas[1], Maxwell B. Colonna[2]

*Corresponding author: ricardo_mborges@ufrj.br

[1]Instituto de Pesquisas de Produtos Naturais Walter Mors, Universidade Federal do Rio de Janeiro, Brazil.

[2]Departments of Genetics and Biochemistry & Molecular Biology, Complex Carbohydrate Research Center, University of Georgia, USA.

Graphical Abstract



Abstract

Natural products and metabolomics are intrinsically linked by the efforts of analyzing complex mixtures for compound annotation. Although most of the studies that aims for compound identification in mixtures use MS as the main analysis technique, NMR has complementary advances that are worth exploring for enhanced structure confidence. This review intends to showcase a portfolio of the main tools available for compound identification using NMR. COLMAR, SMART-NMR, MADByTE, and NMRfilter are presented using examples collected with real samples from the perspective of a natural products chemist.

1. INTRODUCTION

A growing concern in natural products, food and drug discovery research is the progressive loss of biodiversity and the extinction of unexplored species. This gradual extinction of many unexplored plants and other organisms represents the loss of many potentially new bioactive and valuable chemicals.[1] Thus, there is a need for new approaches for accurate documentation of these compounds to allow an efficient cataloging of natural products. Mainly due to the wide range of structural diversity, phytochemistry has become a science focused on progressive fractionations to a pure compound to have its structure and biological activity elucidated. However, the idea of chance once associated with natural products discoveries is now outdated and this major field of research is becoming driven by the 'omics' technologies.[2] Genomics, transcriptomics, proteomics, and metabolomics form a holistic approach to a better understanding of complex biological systems. The first three will not be addressed directly in here. Metabolomics, as the last of the 'omics' sciences, is the term used to describe an analytical approach that aims to identify and quantify all metabolites involved in a specific situation within a given organism or system. It uses analytical measurements and statistical techniques to deliver phenotypic results to be interpreted as the ultimate consequence of genome and proteome modulation.[3] Once there is statistical consideration, a sizeable number of replicate samples must be analyzed to provide statistical confidence.[4]

Considering both natural products and metabolomics (in its wide application), one of the main bottlenecks is compound identification. The term "metabolomics dark matter"[5] was coined to emphasize this critical fact that most compounds in a study are actually unidentified. When the selected biomarker is known and its spectrum is recorded in an available database, one can simply compare the analytical data and identify it. This is routinely done for biological samples, especially in studies involving metabolomics in humans (and other model organisms), where the wide range of compounds is well known and well recorded in various databases. This is not quite the same for natural products, where the chemical diversity is much wider, with varying physicochemical properties, and the available databases are not well-organized, comprehensive, or freely accessed in many cases. The complex biosynthesis of secondary metabolites leads to the possibility of finding many analog derivatives at different stages of biosynthesis, as well as its metabolized forms. In such cases, many of the diagnostic peaks will overlap, causing misunderstandings between multiple possible targets. In natural products, the term dereplication is defined as the approach that allows the identification of known compounds in the early stages of research, in order to avoid investments in the production of known (or replicated) results.[6] This approach relies heavily on databases for matching and identifying chemical profiles in often complex samples. However, because databases are unable to represent all known compounds, we can redefine dereplication as the effort to identify known, well-cataloged compounds in early stages of the research, which is not enough.

Mass Spectrometry (MS) and Nuclear Magnetic Resonance (NMR) are the main techniques for studies at the interface of life sciences and chemistry. MS is mostly used due to its high sensitivity and wider spectral database, but the need for a specific standard for quantitative results and the intrinsic deficiency to provide unambiguous structural identification are limiting for its sole use for analysis. On the other hand, NMR is the most structurally informative analytical tool in organic chemistry and its capability as a direct quantitative detection brands it as an absolute detector. The complementary aspects of NMR and MS are obvious, and NMR disadvantages are related to its low sensitivity and poorly cataloged databases. The complexity presented by NMR spectra, resulting from the fact that a molecule produces more than one signal, can be reflected as an advantage, since these signals characterize

structural particularities for the unequivocal identification of organic compounds. A wide variety of small molecule metabolites can be measured simultaneously in NMR. In fact, the only requirement for NMR signal detection is the presence of active nuclei (e.g. the ^{1}H isotope naturally occurring in 99.97%); solvent solubility for liquid-state NMR are also important requirements for high quality data. But the intense overlap of signals that occurs in complex mixtures in the NMR spectra can limit a complete qualification of the analyzed sample since each compound can present a very complex spectrum by itself. With the use of two-dimensional (2D) experiments, mainly those modulated by the chemical shift of ^{13}C, there is an increase in the spectral window to 20 x 200 ppm which makes the overlap less likely, compared to the spectral window of 20 ppm in an NMR experiment of one-dimensional ^{1}H. In this sense, with the use of 2D experiments, in principle, it is possible to obtain chemical shift data of two interconnected nuclei for comparison with known database for matching, while, in one-dimensional (1D) experiments, the comparison is made using chemical shift data of only one type of nucleus. Similarly to the use of MS/MS data to assure compound annotation in MS, the orthogonal dimension in NMR can also be used to assure identification to organic compounds.

This review aims to explore the main tools that are freely available for compound identification in mixtures using 2D NMR and to suggest a new pulse sequence that could be used for time-saving acquiring the spectral data required by these methods. This is meant to be a guide for research groups in the field of natural products and metabolomics to explore the available possibilities and the used data is accessible here (ZENODO at DOI: https://doi.org/10.5281/zenodo.7601517). The demonstrations that follow were done using data from open repository and data collected from samples processed by the authors. Method descriptions for sample preparation are available at the Supplementary Information (SI.1)

2. MAIN TOOLS AVAILABLE FOR COMPOUND IDENTIFICATION USING 2D NMR

2.1. COLMAR

The Complex Mixture Analysis by NMR (COLMAR; http://spin.ccic.ohio-state.edu/index.php/colmar, coordinated by Prof. Rafael Brüschweiler at The Ohio State University)[7-9] is the most used method for compound identification in mixtures with NMR data. It is well established mainly within the community involved in NMR metabolomics, as this method is accessible via a web server where different analyses can be explored. Importantly, COLMAR has unified metabolite data from two of the most important databases available: BMRB (Biological Magnetic Resonance Data Bank) and HMDB (Human Metabolome Database). Recently, nmrshiftdb2 was included for library matching using $CDCl_3$ for lipids identification. This makes it one of the most comprehensive tools for matching experimental NMR data of complex mixtures to NMR databases. Among the most used features, users can submit their data (^{1}H NMR, ^{13}C NMR, HSQC and/or TOCSY) for a sequence of library comparisons with data from spectral libraries available in their structure. Thus, users can interactively detect which peaks (and how many peaks) of a particular compound exist in their data and judge their presence within the sample. Although this is a useful tool for analyzing primary metabolites, it may be limited to studies with secondary metabolites. COLMAR relies on high quality experimental spectra acquired using standard conditions and the most commonly used solvent for this database is D_2O-phosphate buffer at pH 7.4, although it has been recently complemented by a library of less hydrophilic compounds analyzed in $CDCl_3$. Importantly, matching of experimental data to database are dependent on accurate chemical shift, whereas the chemical shift of ^{1}H-NMR are very dependent on solvation, solvent pH and

temperature. Many compounds of interest in natural products, however, are not soluble in D_2O or $CDCl_3$ and their data are also acquired using CD_3OD or DMSO.*d*6, which makes direct comparison with the experimental database problematic.

Briefly, in the COLMARm method, user can submit experimental ^{13}C- ^{1}H HSQC, ^{1}H-^{1}H TOCSY and ^{13}C-^{1}H-^{1}H HSQC-TOCSY data to perform high confidence searches and identification. First, a computational matching of the experimental HSQC data to database spectra leads to a visualization of matched peaks and a list of candidates is presented. Secondly, the user will be able to validate the identification made with the HSQC using the (HSQC-)TOCSY data that lead to a greater selectivity due to the exploitation of the information transferred via chemical bond within the $^{1}H/^{13}C$-^{1}H spin system, in a scalar coupling ($^{2,3}J_{HH}$). Spectra can be processed using different software and exported into a file that will be readable by COLMAR (e.g. MNOVA's csv files and NMRPipe's ft files). The importance of COLMAR to the scientific community is reflected in more than 70 citations and more than 2000 registered views. For studies focusing on primary metabolites, COLMAR is the tool of choice if samples are water-soluble.

Figure 1 shows the resulting plot produced by the method COLMARm. The resulting compound report list produced using COLMARm for the HSQC and TOCSY data from the raw methanol-water extract of *Caenorhabditis elegans* is presented in full in the SI (**Table SI.2.1** and **Figure SI.2.1**). Another example using a raw methanol-water *Ginkgo biloba* extract is shown in the SI to provide a demonstration of a known natural product sample in both D_2O-phosphate buffer at pH 7.4 and DMSO.*d*6. The resulting compound report lists produced using COLMARm for the HSQC and TOCSY data from the *G. biloba* extract in both deuterated solvents are presented in full (**Table SI.2.2** and **Table SI.2.3**).

As expected, more compounds were annotated and validated using the additional spin system information provided by the TOCSY data from the *C. elegans* data than the *G. biloba* data in both solvents. This can be explained by the fact that the extracts of *G. biloba* is rich in secondary metabolites that outweigh the amount of primary metabolites. BMRB and HMDB contains mostly primary metabolites which can be readily identified in biofluids NMR data. No identification was achieved using the data collected from the *G. biloba* samples. Besides many compounds were annotated for the *G. biloba* HSQC data, the validation using the spin system provided by the TOCSY data failed to confirm any of the suggestions made.

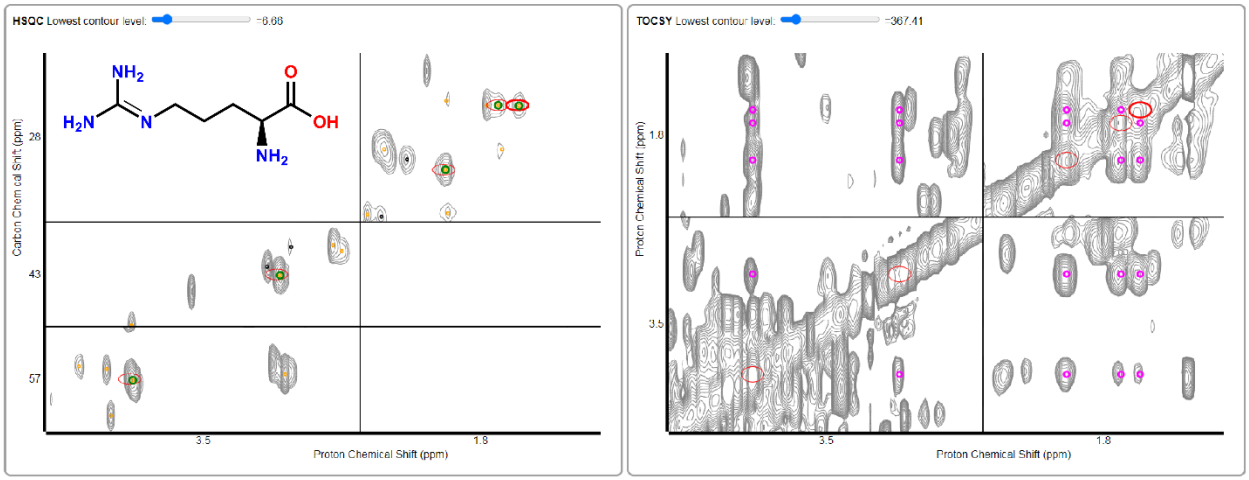


Figure 1. Example of the visual validation of L-arginine from the *C. elegans* sample performed using COLMARm (https://spin.ccic.osu.edu/index.php/colmarm).

2.2. SMART-NMR

The Small Molecule Accurate Recognition Technology (SMART 2.0; https://smart.ucsd.edu/classic, coordinated by Prof. William Gerwick at the University of California San Diego),[10] has a different but complementary approach. It is an infrastructure that uses machine learning (ML) with an artificial intelligence (AI) algorithm based on convolutional neural networks (CNN) to allocate the experimental HSQC data in a multidimensional space (the moliverse). The spacial position of the queried data in the moliverse suggests, by approximation, the identity of the compounds occuring in the sample, promising to accelerate the rate of discovery of new natural products. The idea is based on a concept similar to that used in facial recognition, where the HSQC data replace the photos. The moliverse is composed of over 100,000 known compounds with both experimental and simulated NMR data. In the SMART web portal the experiments can be performed and the results are shown in the form of SMILES structures (representation of chemical structures using normal characters; Simplified Molecular-Input Line-Entry System) together with the similarity parameters (cosine), molecular mass and links for interaction with other databases, namely GNPS,[11] MIBiG[12] e NPAtlas.[13]

The application of SMART-NMR for mixtures is limited since it depends on the peak profile in the HSQC spectra. It can alocate a pure spectra within the chemical space of its molecular structure quite successfully to aid the structure elucidation of purified compounds. In addition, it has already been used for mixtures (fractions of a chromatographic separation) and the results were helpful in prioritizing samples for structure determination. Together with LC-MS/MS and bioactivity data, SMART-NMR enables the early detection of potentially active compounds.[14] This is definetely a tool worth being utilized, especially for those struggling with the assignment of peaks to new structures since it suggests similar structures from the HSQC spectra. A demonstration was done using data of azythromycin, paclitaxel, and ginsenoside RG1. These three datasets were obtained from the fiblib of Agilent/Varian software. **Figure 2** shows the results produced by SMART-NMR for azythromycin. The analyses using SMART-NMR for the HSQC data of azythromycin, paclitaxel, and ginsenoside RG1 are presented in full in SI (**Figures SI.3.1**, **SI.3.2**, and **SI.3.3**, and **Tables SI.3.1**, **SI.3.2**, and **SI.3.3**). When submitting the peak picking data from paclitaxel to SMART-NMR, it promptly suggested “taxol” as the heighest score and the users are directed to the GNPS MSMS-based database to enable a complementary compound identification. The HSQC data from azythromycin was matched to azythromycin within the 7 top scores, together with other close-related structures (e.g.: erythomycin, Iankamycin, clarithromycin). The identity of ginsenoside RG1 was not suggested directly from HSQC data, but it suggested closely related structures that can certainly be helpfull to structure elucidation. Even the early classification of the core structure of complex natural product serves as a great aid in the identification of new compounds. Notably, the suggested candidates and the links to interact with other databases can be usefull for data interpretation.

To prove its usefulness with complex mixture samples, the HSQC data from *G. biloba* was also submited to SMART-NMR (**Table SI.3.4**). More divergent compounds were suggested since more compounds are likely to be present in a raw extract. Since this review does not intend to fully analysis this data, a exhaustive evaluation of the SMART-NMR output will not be done. If that would be the case, a full LC-MSMS dereplication procedure would be valuable to yield a list of possible compounds to be confirmed.

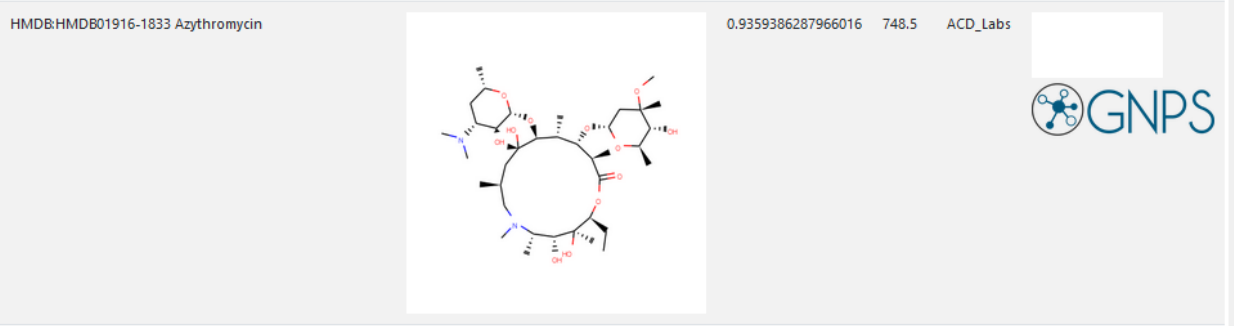


Figure 2. Azythromycin as suggested using SMART-NMR with the HSQC data from azythromycin.

2.3. MADByTE

Metabolomics and Dereplication by Two-Dimensional Experiments[15] (MADByTE; coordinated by Dr. Joseph M. Egan at Simon Fraser University) is a program that allows performing a comparative analysis of experimental HSQC and TOCSY spectra of a set of samples simultaneously. Thus, it allows the visualization of common profile peaks for prioritization of samples and rationalization based on the presented substructures. Specifically, these common features are spin systems identified in the TOCSY data and their respective ^{13}C chemical shifts obtained via the HSQC data indexed by the ^{1}H chemical shift. Then, MADByTE identifies spin systems between different samples and assembles a network composed of nodes, which represent each sample, and edges related to the occurrence of certain spin system in each sample or shared between different samples. In fact, this visualization mode becomes more interesting when different samples are submitted to NMR analysis and to biological activity assays, when it is possible to call attention to the nodes with a promising biological response and observe the spin systems that characterize them. Another interesting application is demonstrated with the use of a database or reference spectra of compounds similar to those expected to be found in the samples to facilitate the identification of shared substructures. **Figure SI.4.1** shows the interface of MADByTE.

The use of MADByTE is limited by the need to acquire both HSQC and TOCSY data for multiple samples (or fractions) of a study. This effort is not routine for natural products chemists due to the cost of long periods of NMR use. The advances of NMR by Ordered Acquisition using ^{1}H detection (NOAH supersequences)[16, 17], which combine different ^{1}H detection 2D NMR experiment into a “supersequence” eliminating the individual relaxation period (d1) of each combined unit, will overcome this limitation. In fact, the experiment time of a combined HSQC and TOCSY under the same NOAH supersequence is comparable to the acquisition of HSQC or TOCSY data alone, considering the same sequence parameters. This approach was demonstrated using the COLMAR platform where HSQC and HSQC-TOCSY data had been acquired in a single pulse sequence indicating a time savings of around 40%.[16] Due to the limitation in collection HSQC and TOCSY data for a batch of samples for this demonstration, an example made available by Egan’s research group was used.[15] The output network is shown at **Figure 3** with nodes to represent samples and co-occurring spin systems.

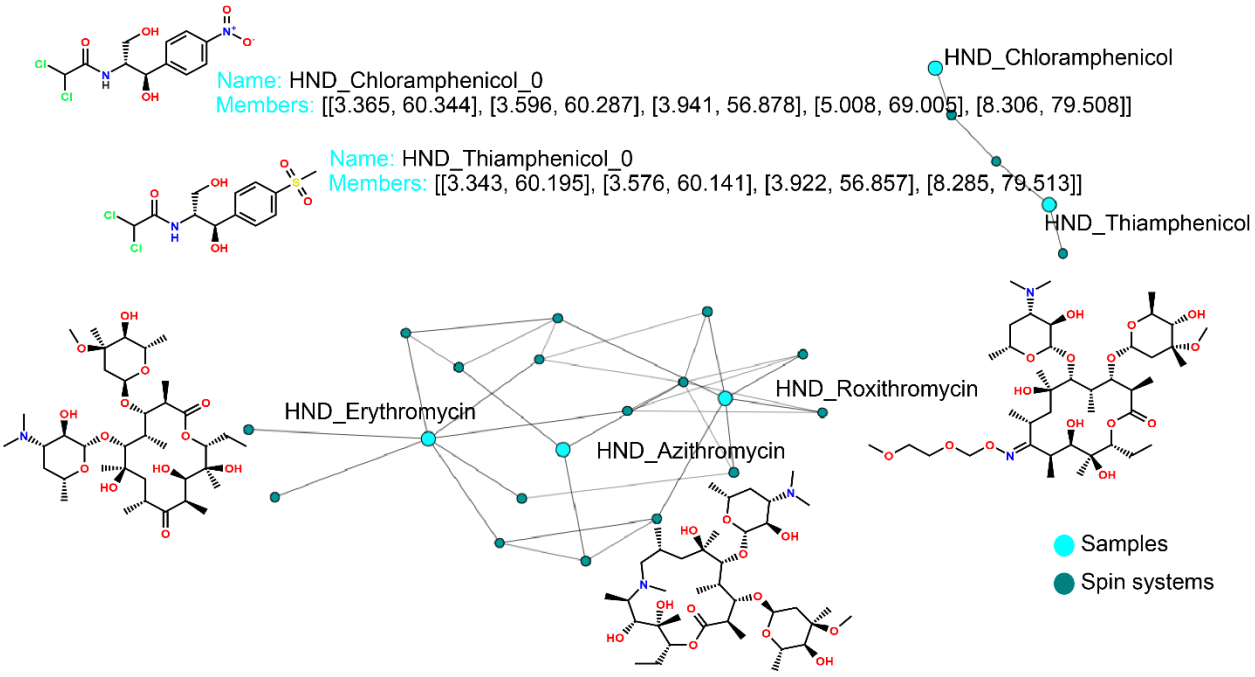


Figure 3. MADByTE network representing nodes (light blue: samples; dark blue: spin systems) and edges to show co-occurrence of similar TOCSY derived spin-systems. Test boxes show the information made available by hovering over the nodes. The TOCSY derived spin systems are detailed on the dark blue nodes together with their respective HSQC derived data.

2.4. NMRfilter

The NMRfilter [6, 18] (https://github.com/stefhk3/nmrfilter; coordinated jointly by Prof. Ricardo M. Borges at the Federal University of Rio de Janeiro and Prof. Stefan Kuhn at University of Tartu) has an approach that originated with the intention of combining the interpretation of MS data with NMR data. The algorithm introduced by this method starts from a list of candidates (compounds possibly existing in the sample) that will have their chemical shift data simulated and compared with the users' experimental data submitted to NMRfilter. Then, a search for heteronuclear and homonuclear scalar coupling networks is performed on HMBC and HSQC-TOCSY data to define whether the signals matched in the HSQC share a unique chemical structure. In this case, it is assumed that, ideally, all nuclei of a compound will correlated with each other via $^{2\text{-}4}J_{CH}$. As with SMART-NMR, the dependence of an experimentally obtained database has a lower weight once the NMR data of the expected compounds can be simulated. However, this simulation is based on data cataloged in nmrshiftdb2 (https://nmrshiftdb.nmr.uni-koeln.de/)[19, 20] and the more complete this data source, the more accurate the simulated data. In other words, more precise NMR data simulations will be achieved when the database possesses similar compounds, especially for the Hierarchical Organization of Spherical Environments (HOSE) like method. [21] Thus, users are invited to submit NMR data from the literature to nmrshiftdb2 to populate the database for better simulations.

Figure 4 shows an example result from NMRfilter for the identification of quercitrin in the raw extract of *G. biloba* in deuterated DMSO. HSQC, HMBC and HSQC-TOCSY data were acquired as

described at SI.1.5. Note the list of candidates suggested and the visual validation where users can evaluate both the matching rates and the peak profiles to assert their identification. Some of resulting figures and table produced by NMRfilter using the *G. biloba* data are available (Table SI.5.1, and Figure SI.5.1). Similarly to COLMARm, which uses the TOCSY spin system for candidate confirmation, NMRfilter uses heteronuclear HMBC spin system (and the TOCSY spin system via HSQC-TOCSY when available) to detect the spin fragments.

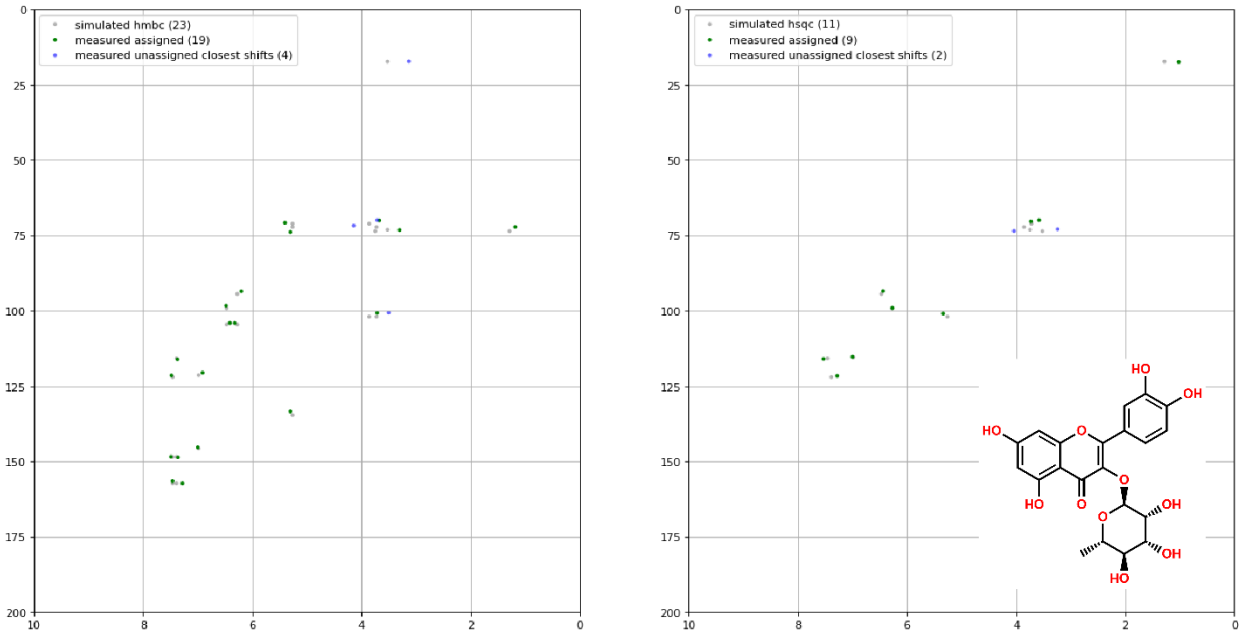


Figure 4. Visual validation of quercitrin (LTS0093095 from the Lotus database at lotus.naturalproducts.net) produced by NMRfilter from the data of the *G. biloba* extract.

NOAH supersequences

It is important to mention the use of NOAH supersequences to acquire different 2D NMR data within the same experiment. The time savings provided by these sequences are a consequence of the use of a single recovery time (parameter *d1* on Bruker spectrometers) for all the modules (individual pulse sequence).[17] These new sequences are receiving considerable attention from the NMR community specially for structure elucidation,[17, 22] but also for metabolomics (complex mixture) samples.[16] A combination of a selective enhanced HSQC together with a TOCSY module yields data useful for library matching and confirmation by spin systems identification (using COLMARm) in around 60% of the total time required for the acquisition of both experiments independently; we can also advocate for the benefits of these from our own experience. This sequence can be downloaded from the genesis web portal (https://nmr-genesis.co.uk/)[23] as gns_noah2-SpT (seHSQC + TOCSY as suggested at [16]). Following the same rational described here, data acquired using the same NOAH supersequence gns_noah2-SpT could also be used for applications to MADByTE. There is virtually no sensitivity loss for using this experiment due to its specific construction and spin dynamics. Briefly, the HSQC data is collected using only the magnetization from ^{1}H directly linked to ^{13}C, while the magnetization from ^{1}H directly linked to ^{12}C (~99%) is kept safe for the TOCSY module to produce comparable data. This isotopic magnetization

selection is enabled by the zz-filter (or ZIP sequence) implemented in the HSQC module. Other combinations are being tested to combine different modules to enable the acquisition of HMBC, HSQC, and TOCSY (or HSQC-TOCSY) data within the same NOAH supersequence.

Considering the experimental requirement for the tools for compound identification presented here, we evaluated the use of a NOAH supersequence comprising of HMBC, HSQC, and TOCSY, named gns_noah3-BSpT downloaded from the genesis web portal.[23] A comparison between the HSQC spectra acquired from hsqcedetgpsis2 and gns_noah3-BSpT is shown (Figure 5). Rich spectra are shown for both experiments as it was expected, but the spectra acquire using gns_noah3-BSpT has shown more peaks and higher intensity peaks than the conventional hsqcedetgpsis2. A full comparison between these experiments is outside the scope of this review, but NOAH supersequences are strongly encouraged not only for pure compounds and structure elucidation to save NMR instrument time.[17, 22] The COLMARm and MADByTE uses HSQC and TOCSY (and/or HSQC-TOCSY) data, SMART-NMR uses HSQC data, and NMRfilter uses mainly HMBC and HSQC data (HSQC-TOCSY data can be used but it is not required). This evaluation is presented at figure SI.6.1-SI.6.3 using the *G. biloba* extract sample. Noticeably, the time savings for the acquisition of gns_noah3-BSpT using the same parameters was found to be around 45% (12 h 18 min for 32 scans and 1.5 s of recovery time) compared to the use of hmbcetgpl3nd (HMBC), hsqcetgpprsisp2.3 (HSQC), and dipsi2gpphpr (TOCSY) together; Non-Uniform Sampling (NUS) is also possible to be used with these options. This data set can then be used with all the aforementioned tools for better coverage of annotations. The idea is to acquire enough data to enable users to use the plethora of tools to aid compound annotation in early stages of chemical profiling studies.

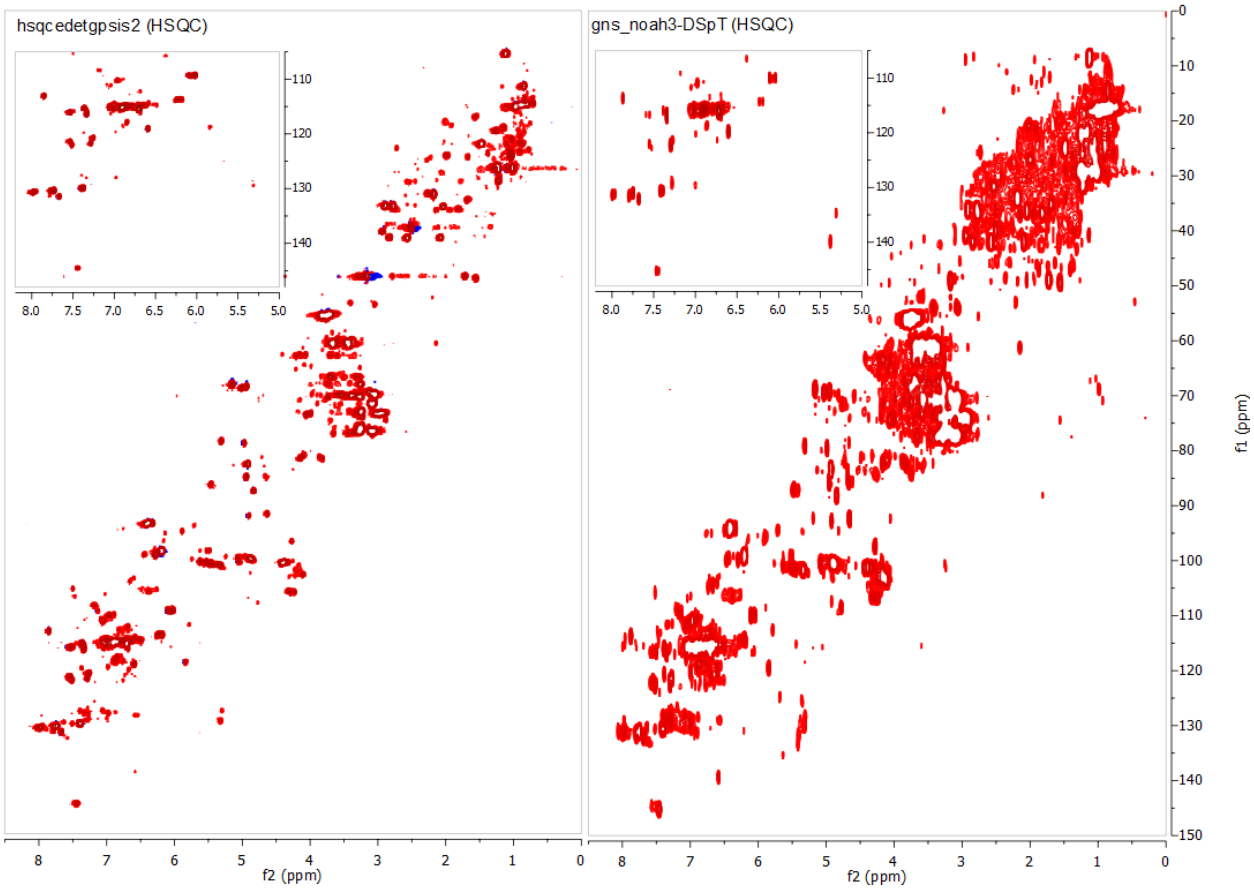


Figure 5. Comparison between the HSQC spectra acquired using hsqcedetgpsis2 and gns_noah3-BSpT from the extract of *G. biloba* in DMSO.d_6. Data collected at a 800 MHz Bruker Avance III equipped with a 1.7 mm TCI cryoprobe.

FINAL CONSIDERATIONS

The acquisition of ^{1}H NMR data, HSQC, TOCSY, HSQC-TOCSY and HMBC for all samples processed, from the initial extract, in the case of a study on natural products, is strongly suggested. The same is valid for MS analysis in order to create a bridged analysis between both techniques in a complementary manner. Understanding that this is, perhaps, an unreasonable expectation due to the time (and costs) required to acquire each experiment, the use of NOAH supersequences are strongly suggested.[17] This recommendation was highlighted by a recent publication proving that NOAH supersequences can be successfully used to collect data for COLMARm to achieve library matching. NUS, which was not detailed here, is also suggested whenever possible when there is enough sample for adequate sensitivity.

Every year new processing methods have been developed and made available to the scientific community both in the field of metabolomics and in natural products.[6, 24-27] The goals are, generally, to facilitate the process of annotation of constituent compounds in samples and to aid decision-making

towards new research. In addition, the application of methods to support in the endeavor to catalog the biodiversity that has been continuously lost should be highly valued.

ACKNOWLEDGMENT

The authors acknowledge Walter Mors Institute of Research on Natural Products, Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) for the grant 210.489/2019 APQ-1. The Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) Brazilian Federal Funding Agency Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) are also acknowledged.

REFERENCES

1. Newman DJ and Cragg GM. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *Journal of Natural Products*. 2020;83:770-803.
2. Wolfender JL, Litaudon M, Touboul D and Queiroz EF. Innovative omics-based approaches for prioritisation and targeted isolation of natural products - new strategies for drug discovery. *Nat Prod Rep*. 2019;36:855-868.
3. Sumner LW, Mendes P and Dixon RA. Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry*. 2003;62:817-836.
4. Blaise BJ, Correia G, Tin A, Young JH, Vergnaud AC, Lewis M, Pearce JT, Elliott P, Nicholson JK, Holmes E and Ebbels TM. Power Analysis and Sample Size Determination in Metabolic Phenotyping. *Anal Chem*. 2016;88:5179-5188.
5. Peisl BYL, Schymanski EL and Wilmes P. Dark matter in host-microbiome metabolomics: Tackling the unknowns-A review. *Anal Chim Acta*. 2018;1037:13-27.
6. Kuhn S, Colreavy-Donnelly S, Santana de Souza J and Borges RM. An integrated approach for mixture analysis using MS and NMR techniques. *Faraday Discuss*. 2019;218:339-353.
7. Bingol K, Li DW, Zhang B and Bruschweiler 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.
8. Wang C, Timari I, Zhang B, Li DW, Leggett A, Amer AO, Bruschweiler-Li L, Kopec RE and Bruschweiler R. COLMAR Lipids Web Server and Ultrahigh-Resolution Methods for Two-Dimensional Nuclear Magnetic Resonance- and Mass Spectrometry-Based Lipidomics. *J Proteome Res*. 2020;19:1674-1683.
9. Bingol K, Li D-W, Bruschweiler-Li L, Cabrera OA, Megraw T, Zhang F and Brüschweiler R. Unified and Isomer-Specific NMR Metabolomics Database for the Accurate Analysis of 13C–1H HSQC Spectra. *ACS Chemical Biology*. 2015;10:452-459.
10. Reher R, Kim HW, Zhang C, Mao HH, Wang M, Nothias LF, Caraballo-Rodriguez AM, Glukhov E, Teke B, Leao T, Alexander KL, Duggan BM, Van Everbroeck EL, Dorrestein PC, Cottrell GW and Gerwick WH. A Convolutional Neural Network-Based Approach for the Rapid Annotation of Molecularly Diverse Natural Products. *J Am Chem Soc*. 2020;142:4114-4120.
11. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu WT, Crusemann M, Boudreau PD, Esquenazi E, Sandoval-Calderon M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu CC, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw CC, Yang YL, Humpf

HU, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, P CAB, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai J, Neupane R, Gurr J, Rodriguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard PM, Phapale P, Nothias LF, Alexandrov T, Litaudon M, Wolfender JL, Kyle JE, Metz TO, Peryea T, Nguyen DT, VanLeer D, Shinn P, Jadhav A, Muller R, Waters KM, Shi W, Liu X, Zhang L, Knight R, Jensen PR, Palsson BO, Pogliano K, Linington RG, Gutierrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC and Bandeira N. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol*. 2016;34:828-837.

12. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, van Santen JA, Tracanna V, Suarez Duran HG, Pascal Andreu V, Selem-Mojica N, Alanjary M, Robinson SL, Lund G, Epstein SC, Sisto AC, Charkoudian LK, Collemare J, Linington RG, Weber T and Medema MH. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Research*. 2019;48:D454-D458.

13. van Santen JA, Jacob G, Singh AL, Aniebok V, Balunas MJ, Bunsko D, Neto FC, Castano-Espriu L, Chang C, Clark TN, Cleary Little JL, Delgadillo DA, Dorrestein PC, Duncan KR, Egan JM, Galey MM, Haeckl FPJ, Hua A, Hughes AH, Iskakova D, Khadilkar A, Lee JH, Lee S, LeGrow N, Liu DY, Macho JM, McCaughey CS, Medema MH, Neupane RP, O'Donnell TJ, Paula JS, Sanchez LM, Shaikh AF, Soldatou S, Terlouw BR, Tran TA, Valentine M, van der Hooft JJJ, Vo DA, Wang M, Wilson D, Zink KE and Linington RG. The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. *ACS Cent Sci*. 2019;5:1824-1833.

14. Kim HW, Kim SS, Kang KB, Ryu B, Park E, Huh J, Jeon WK, Chae HS, Oh WK, Kim J, Sung SH and Chin YW. Combined MS/MS-NMR Annotation Guided Discovery of Iris lactea var. chinensis Seed as a Source of Viral Neuraminidase Inhibitory Polyphenols. *Molecules*. 2020;25.

15. Egan JM, van Santen JA, Liu DY and Linington RG. Development of an NMR-Based Platform for the Direct Structural Annotation of Complex Natural Products Mixtures. *J Nat Prod*. 2021.

16. Hansen AL, Kupce ER, Li DW, Bruschweiler-Li L, Wang C and Bruschweiler R. 2D NMR-Based Metabolomics with HSQC/TOCSY NOAH Supersequences. *Anal Chem*. 2021;93:6112-6119.

17. Kupce E and Claridge TDW. NOAH: NMR Supersequences for Small Molecule Analysis and Structure Elucidation. *Angew Chem Int Ed Engl*. 2017;56:11779-11783.

18. Kuhn S, Colreavy-Donnelly S, de Andrade Silva Quaresma LE, de Andrade Silva Quaresma E and Borges RM. Applying NMR compound identification using NMRfilter to match predicted to experimental data. *Metabolomics*. 2020;16:123.

19. Kuhn S, Egert B, Neumann S and Steinbeck C. Building blocks for automated elucidation of metabolites: machine learning methods for NMR prediction. *BMC Bioinformatics*. 2008;9:400.

20. Steinbeck C and Kuhn S. NMRShiftDB -- compound identification and structure elucidation support through a free community-built web database. *Phytochemistry*. 2004;65:2711-2717.

21. Kuhn S and Johnson SR. Stereo-Aware Extension of HOSE Codes. *ACS Omega*. 2019;4:7323-7329.

22. Claridge TDW, Mayzel M and Kupce E. Triplet NOAH supersequences optimised for small molecule structure characterisation. *Magn Reson Chem*. 2019;57:946-952.

23. Yong JRJ, Kupce ER and Claridge TDW. Modular Pulse Program Generation for NMR Supersequences. *Anal Chem*. 2022;94:2271-2278.

24. Borges RM, das Neves Costa F, Chagas FO, Teixeira AM, Yoon J, Weiss MB, Crnkovic CM, Pilon AC, Garrido BC, Betancur LA, Forero AM, Castellanos L, Ramos FA, Pupo MT and Kuhn S. Data Fusion-based Discovery (DAFdiscovery) pipeline to aid compound annotation and bioactive compound discovery across diverse spectral data. *Phytochem Anal*. 2022.

25. Rogers S, Ong CW, Wandy J, Ernst M, Ridder L and van der Hooft JJJ. Deciphering complex metabolite mixtures by unsupervised and supervised substructure discovery and semi-automated annotation from MS/MS spectra. *Faraday Discuss*. 2019;218:284-302.
26. Nothias LF, Petras D, Schmid R, Duhrkop K, Rainer J, Sarvepalli A, Protsyuk I, Ernst M, Tsugawa H, Fleischauer M, Aicheler F, Aksenov AA, Alka O, Allard PM, Barsch A, Cachet X, Caraballo-Rodriguez AM, Da Silva RR, Dang T, Garg N, Gauglitz JM, Gurevich A, Isaac G, Jarmusch AK, Kamenik Z, Kang KB, Kessler N, Koester I, Korf A, Le Gouellec A, Ludwig M, Martin HC, McCall LI, McSayles J, Meyer SW, Mohimani H, Morsy M, Moyne O, Neumann S, Neuweger H, Nguyen NH, Nothias-Esposito M, Paolini J, Phelan VV, Pluskal T, Quinn RA, Rogers S, Shrestha B, Tripathi A, van der Hooft JJJ, Vargas F, Weldon KC, Witting M, Yang H, Zhang Z, Zubeil F, Kohlbacher O, Bocker S, Alexandrov T, Bandeira N, Wang M and Dorrestein PC. Feature-based molecular networking in the GNPS analysis environment. *Nat Methods*. 2020;17:905–908.
27. Kuhn S, Cobas C, Barba A, Colreavy-Donnelly S, Caraffini F and Borges RM. Direct deduction of chemical class from NMR spectra. *J Magn Reson*. 2023;348:107381.