Metabolomics (2020) 16:123
https://doi.org/10.1007/s11306-020-01748-1

SHORT COMMUNICATION

Applying NMR compound identification using NMRfilter to match predicted to experimental data

Stefan Kuhn[1] · **Simon Colreavy-Donnelly**[1] · **Lucas Eliseu de Andrade Silva Quaresma**[2] · **Ezequiel de Andrade Silva Quaresma**[2] · **Ricardo Moreira Borges**[2]

Received: 11 May 2020 / Accepted: 11 November 2020 / Published online: 21 November 2020

Abstract

Introduction Metabolomics is the approach of choice to guide the understanding of biological systems and its molecular intricacies, but compound identification is yet a bottleneck to be overcome.

Objective To assay the use of NMRfilter for confidence compound identification based on chemical shift predictions for different datasets.

Results We found comparable results using the lead tool COLMAR and NMRfilter. Then, we successfully assayed the use of HMBC to add confidence to the identified compounds.

Conclusions NMRfilter is currently under development to become a stand-alone interactive software for high-confidence NMR compound identification and this communication gathers part of its application capabilities.

Keywords NMR · Compound identification · Metabolomics · Dereplication · NMRfilter

1 Introduction

Today, much of the effort in life science can be summarized as understanding biological systems, and so, metabolomics has emerged as the approach of choice. From fundamental ecological and interaction studies to precision medicine, metabolomics has been applied with promising results due to its scrutiny to establish statistically supported biomarkers when different groups are compared (Wishart 2019). Although much was accomplished in experimental design, mathematical modeling, and statistical protocols, one major bottleneck is yet to be solved: Unequivocal compound identification. Natural product research is another major area where this is important (Hubert et al. 2017).

When dealing with samples consisting mainly of primary metabolites, such as in biofluids, methods for compound identification based on formal databases are straightforward. Complex Mixture Analysis by NMR (COLMAR; http://spin.ccic.ohio-state.edu/index.php/colmar) (Bingol et al. 2015) is a leading system that runs a matching algorithm for chemical shift comparison using the Biological Magnetic Resonance Data Bank (BMRB) and the Human Metabolome Database (HMDB). COLMAR is successfully and broadly used across the literature for compound identification yielding confidence parameters. The one drawback of such database-driven methods is its strong reliance on how comprehensive those databases are.

A valid alternative to access NMR data from non-cataloged (and even for unknown) compounds is to make use of predictive methods. Our group previously reported a method that integrates the results of an MS-driven dereplication into an NMR peak matching routine (Kuhn et al. 2019). NMRfilter is part of this algorithm that runs an NMR chemical shift predictions and matches them with the experimental data. Users would then define the identity of such compounds using a list of matching rates and correlating parameters of accuracy together with figures for visual validation.

The strategy followed here was as follows. Firstly, we validate the use of NMRfilter as a valid identification

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11306-020-01748-1) contains supplementary material, which is available to authorized users.

✉ Ricardo Moreira Borges
ricardo_mborges@ufrj.br

[1] School of Computer Science and Informatics, De Montfort University, The Gateway, Leicester LE1 9BH, UK

[2] Walter Mors Institute of Research on Natural Products, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

routine for NMR data of mixtures based on the compound list retrieved from COLMAR. COLMAR is the current technique of choice and it is fully dependent on high-quality experimental databases. Thus, we use strictly the compounds appointed by COLMAR as the list of candidates to access the HSQC data only. Afterwards, we expand the confidence for each identification by using the available HMBC data. Through this, we intend to prove the value of (1) the predictive tool for uncatalogued compounds, and (2) $^{2,3}J_{CH}$ HMBC spectra to assert peaks connected through bond interactions.

2 Methods

In this, we use data from an artificial mixture, Drosophila, human urine, and plasma. The artificial mixture was achieved by overlaying available data from the BMRB database of flavone, khellin, tropine, quinidine, beta-carotene, cholecalciferol, and 4-isopropylbenzyl alcohol. We focused on data collected in chloroform-*d*1; only the last compound was added with data collected in D_2O. Note that NMRfilter does not include BMRB as a database for the prediction step yet. The Drosophila HSQC peaklist was copied from the COLMAR web server used as a training example for users (HSQC only).

The urine sample was thawed and an aliquot of 400 uL of the centrifuged supernatant was mixed with 200 uL of the phosphate buffer in D_2O at pH 7.4. The plasma sample was thawed and an aliquot of 200 µL mixed with 400 µL of the phosphate buffer in D_2O at pH 7.4. After centrifugation, 500 µL of supernatant of both the urine and the plasma samples were transferred to 5-mm NMR tubes for analysis.

The experimental NMR data for the human urine and plasma were collected using a 600 MHz Bruker Avance III equipped with a 5 mm TCI cryoprobe. The pulse sequence hsqcedetgpsisp2.2 under non-uniform sampling mode (30% of NUS amount and 307 NUS points; 1024 and 2048 points for F2 and F1, respectively) was used to acquire the edited HSQC data (32 scans). The pulse sequence hmbcetgpl3nd under non-uniform sampling mode (30% of NUS amount and 307 NUS points; 1024 and 2048 points for F2 and F1, respectively) was used to acquire the HMBC data (32 scans). All the HSQC and HMBC data collected was processed accordingly and peakpicked. The peaklists were submitted to COLMAR using the HSQC query for compound matching, using 0.03 ppm threshold for ^{1}H chemical shift and 0.3 ppm threshold for ^{13}C chemical shift. The compounds identified by COLMAR were used as the candidate list for NMRfilter, using the same threshold for chemical shift. For the NMRfilter routine, we set the same cutoff for ^{1}H and ^{13}C. Thus, we included the analysis of the HMBC data within NMRfilter for the network analysis. For the analysis done by this study, we considered only the matching rate of over 50%, reflecting identifications where at least 50% of the peaks were found.

The shift prediction is done using data from nmrshiftdb2 and an extended HOSE code algorithm, which respects stereo-chemical configurations (Kuhn and Johnson 2019).

3 Results and discussion

To assay the NMRfilter method as a valid tool to predict and match compounds from a candidate list within the artificial mixture, the NMR peak lists of Drosophila, urine, and plasma were submitted to both COLMAR and NMRfilter under the same threshold for chemical shifts for ^{1}H and ^{13}C (Supplementary Tables S1, S2, S3 and S4). For this initial assay, only the HSQC dataset was used since the goal was to evaluate NMRfilter as a valid chemical shift predictive tool of known compounds, and not to compare both methods. Thus, the compound lists acquired from the COLMAR matching routine for each dataset were used as candidate lists for the NMRfilter routine for the respective analysis.

First, the artificial sample constructed using the NMR data of 7 randomly chosen pure compounds was processed. The NMRfiler result enabled the identification of them all with over three quarters of the peaks identified within the cutoffs (Table 1; *HSQC matching rate column). Note that NMRfilter does not include BMRB as a database for the prediction step yet. The prediction method used in NMRfilter relies on finding atoms with a similar environment and uses

Table 1 NMRfilter resulting list from the artificial sample

Compound name	Distance	Standard deviation	HMBC matching rate	HSQC matching rate
4-isopropylbenzyl alcohol	0.05	0.53	50.00%	100.00%
Tropine	0	0.37	100.00%	100.00%
Beta-carotene	0	1	97.96%	92.86%
Flavone	0	0.71	100.00%	75.00%
Khellin	0.02	0.63	93.33%	83.33%
Quinidine	0.02	0.83	85.96%	81.25%
Cholecalciferol	0.02	0.81	63.10%	81.82%

their shift as prediction. Those, depending on the contents of the database used, might therefore not be 100% accurate.

The Drosophila dataset submitted to COLMAR resulted in a total of 33 identified compounds where 16 of them were shown to have a matching rate of 100% of the peaks (Fig. 1a) and 29 had over 50% of matching rate (Fig. 1b). Considering the NMRfilter results, 9 compounds were shown to have a matching rate of 100% (Fig. 1a) and 28 compounds, 50% of the peaks (Fig. 1b). The urine dataset submitted to COLMAR resulted in a total of 35 identified compounds where 20 of them were shown to have a matching rate of 100% (Fig. 1c) of the peaks and 25 had over 50% of matching rate (Fig. 1d). Considering the NMRfilter results, 14 compounds were shown to have a matching rate of 100% (Fig. 1c) and 29 compounds 50% of the peaks (Fig. 1d). Finally, the plasma dataset submitted to COLMAR resulted in a total of 17 identified compounds where 13 of them were shown to have a matching rate of 100% (Fig. 1e) of the peaks and 17 had over 50% of matching rate (Fig. 1a). Considering the NMRfilter results, 6 compounds were shown to have a matching rate of 100% (Fig. 1e) and 16 compounds 50% of the peaks (Fig. 1b).

Thus, the chemical shift prediction and matching capabilities of NMRfilter have been validated and shown worthy

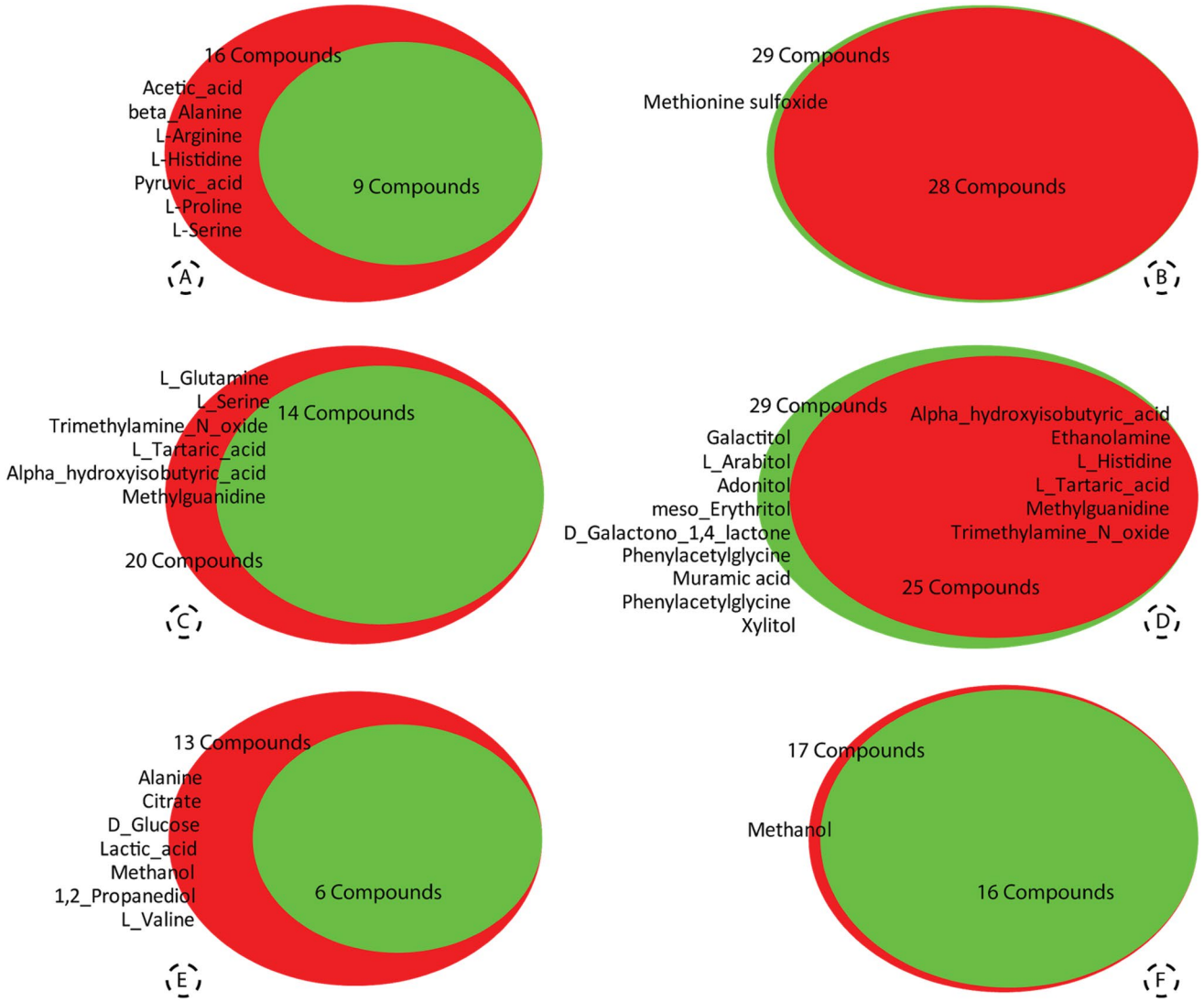

Fig. 1 Comparison between identified compounds using COLMAR (red) and NMRfilter (green) and those compounds identified exclusively by COLMAR or NMRfilter. **a**, **c** and **d** Comprises compounds with all expected peaks matching; 100% matching rate. **b**, **d** and **f** Comprises compounds with half of all expected peaks matching; 50% matching rate. **a** and **b** Results concerning the Drosophila dataset. **c** and **d** Results concerning the urine dataset. **e** and **f** Results concerning the plasma dataset. Note that the candidate list submitted to NMRfilter was the full list appointed by COLMAR using 0.03 ppm threshold for 1H chemical shift and 0.3 ppm threshold for 13C chemical shift

to be applied for a range of sample sources. In this next step, we assay NMRfilter's use to increase the confidence in the identified compounds using the $^3J_{CH}$ based correlation from HMBC. The goal is to show that the matched peaks from the HSQC are in fact connected indicating they share the same chemical structure. The expected drawback in the current dataset is the inherent low intensity of the HMBC peaks.

Including the HMBC into the analysis enables the formation of peak networks across spectra, increasing the chance of the network of a compound to be separable. The parameter 'standard deviation' indicates how much the predicted network matches a measured network.

With the inclusion of the HMBC from the artificial dataset, we added confidence for the identifications. Noteworthy, the high match rate for the HMBC data indicate the confidence added by the method (Table 1). The standard deviation parameters show that the assigned peaks are connected among themselves through bond interactions, and so, they are part of the same compound. Note that we do not mean to present a definitive answer on the mixture composition, but to enable users to gather information to make a data-driven decision. Then, the visual validation step enables by figures should play an important role for the user's decisions (Fig. 2).

The urine dataset confirmed the identification of 10 compounds (3-Hydroxyisovaleric acid, 1-Dimethylbiguanide, Creatinine, Muramic acid, Lactose, Guanidineacetic acid, L-Serine, D-Galactono 1,4-lactone, and L-Histidine) using a 50% threshold for standard deviation with the available HMBC (which had low signal to noise). For the plasma sample, NMRfilter using the HMBC data enabled the confirmation of 9 compounds (L-Proline, L-Valine, L-Glutamine, Lactic-acid, D-Glucose, D-Glucose, 1,2-Propanediol, Taurine, and Leucine) using a 50% threshold for standard deviation.

The key argument for the use of NMRfilter for compound identification by chemical shift matching lies in its capability of identifying uncatalogued compounds. By now, researchers are using mostly experimentally collected data from a formal database (e.g. BMRB, HMDB, and nmrshiftdb2), and so, in a practical sense, they are dealing with uncatalogued known compounds the same way they would with unknown new compounds. Through this, we ask NMR users to submit data from pure compounds and their assigned structures into accessible databases, so it can increasingly improve the prediction accuracy. For instance, nmrshiftdb2 (Kuhn and Schlörer 2015). Additionally, we successfully advocated for use of high-quality HMBC data together with the HSQC for accuracy compound identification using NMR. We strongly suggest the use of HSQC-TOCSY as well, and this can be directly added to the NMRfilter's network analysis; we did not collect any HSQC-TOCSY for this demonstration. All data is available upon request.

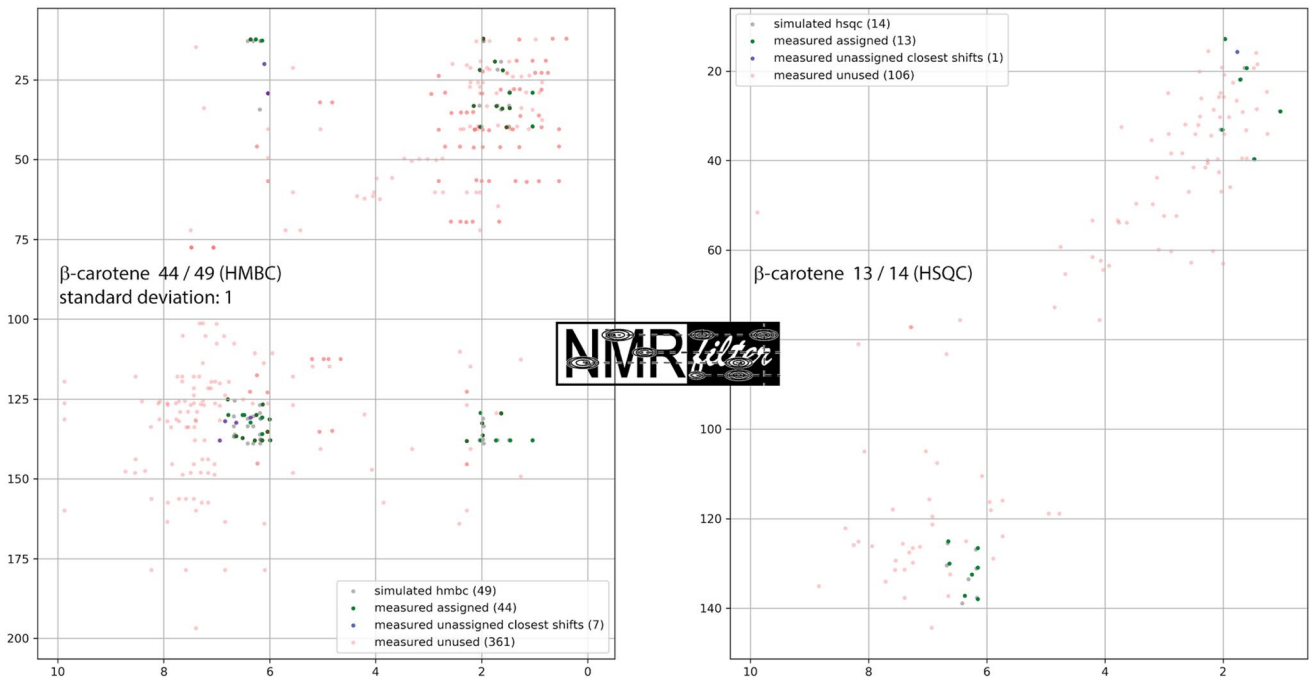

Fig. 2 An example of the visual validation figure created by NMRfilter to access data comparison among the full peak list (in red), the simulated data (in gray) and the matching peaks (assigned peaks in green and closest unassigned peaks in blue). Note that the figure includes the matching rate for each spectrum

Acknowledgements The authors wish to thank the CCRC NMR facility and prof. Dr Arthur S. Edison for the NMR data and Nils Schlörer and the NMR lab at Universität zu Köln for support over many years.

Author contributions SK and RMB conceived this demonstration. SCD conducted part of the programing part of NMRfilter. LWASQ and EZASQ contributed with experiments. All authors read and approved the manuscript.

Data availability The sample data used are available in the github repository at https://github.com/stefhk3/nmrfilterprojects.

References

Bingol, K., Li, D.-W., Bruschweiler-Li, L., Cabrera, O., Megraw, T., Zhang, F., & Brüschweiler, R. (2015). Unified 13C-1H HSQC metabolomics database with isomer-specific query for the analysis of complex metabolite mixtures by NMR. *ACS Chemical Biology, 10,* 452–459.

Hubert, J., Nuzillard, J.-M., & Renault, J.-H. (2017). Dereplication strategies in natural product research: How many tools and methodologies behind the same concept? *Phytochemistry Reviews, 16,* 55–95.

Kuhn, S., Colreavy-Donnelly, S., de Souza, J. S., & Borges, R. M. (2019). An integrated approach for mixture analysis using MS and NMR techniques. *Faraday Discussions, 218,* 339–353.

Kuhn, S., & Johnson, S. R. (2019). Stereo-aware extension of HOSE codes. *ACS Omega, 4*(4), 7323–7329.

Kuhn, S., & Schlörer, N. E. (2015). Facilitating quality control for spectra assignments of small organic molecules: Nmrshiftdb2 – a free in-house NMR database with integrated LIMS for academic service laboratories. *Magnetic Resonance in Chemistry, 53,* 582–589.

Wishart, D. S. (2019). NMR metabolomics: A look ahead. *Journal of Magnetic Resonance, 306,* 155–161.