Journal Name

ARTICLE TYPE

Cite this: DOI: 10.1039/xxxxxxxxxx

An integrated approach for mixture analysis using MS and NMR techniques†

Stefan Kuhn,[a] Simon Colreavy-Donnelly,[a] Juliana Santana de Souza[b] and Ricardo Moreira Borges[b]

Received Date
Accepted Date

DOI: 10.1039/xxxxxxxxxx

+ www.rsc.org/journalname

We suggest an improved software pipeline for mixture analysis. The improvements include combining tandem MS and 2D NMR data for a reliable identification of its constituents in an algorithm based on network analysis aiming for a robust and reliable identification routine. An important part of this pipeline is the use of open-data repositories, although it is not totally reliant on them. The NMR identification step emphasizes robustness and is less sensitive towards changes in data acquisition and processing than existing methods. The process starts with a LC-ESI-MSMS based molecular network dereplication using data from the GNPS collaborative collection. We identify closely related structures by propagating structure elucidation through edges in the network. Those identified compounds are added on top of a candidate list for the following NMR filtering method that predicts HSQC and HMBC NMR data. The similarity of the predicted spectra of the set of closely related structures to the measured spectra of the mixture sample is taken as one indication of the most likely candidates for its compounds. The other indication is the match of the spectra to clusters built by a network analysis from the spectra of the mixture. The sensitivity gap between NMR and MS is anticipated and it will be reflected naturally by the eventual identification of fewer compounds, but with a higher confidence level, after the NMR analysis step. The contributions of the paper are an algorithm combining MS and NMR spectroscopy and a robust $^{n}J_{CH}$ network analysis to explore the complementary aspect of both techniques. This delivers good results even if a perfect computational separation of the compounds in the mixture is not possible. All the scripts will be made available online for users to aid studies such as with plants, marine organisms, and microorganism natural product chemistry and metabolomics as those are the driving force for this project.

1 Introduction

Natural products (NP) are an important source of new pharmacologically active compounds. Regrettably, the rapid extinction of many unexplored plants and other organisms represents losses of a broad range of potential new bioactive and valuable chemicals. An effective and challenge-free method of screening and identifying NP is yet to be well established. Thus, there is a need for new high-throughput approaches to be used as a standard procedure for accurately catalog NP. When a biologically relevant spectral feature is identified and it is listed in a given database, the identification process is straightforward generating a confidence index for the analytical data to a database match. This is routinely done for biological samples especially in human metabolomics studies, where the broad range of compounds is well-known and well-recorded in various databases.[1] This is not the same for NP, where the chemical diversity is much broader with varied physicochemical properties. Their available databases are not well organized, comprehensive or publicly available. The complexity of secondary metabolites biosynthesis leads to the opportunity of uncovering additional compounds at different stages of the biosynthetic/metabolic pathway with similar core structure. Within this context, mass spectrometry (MS) and nuclear magnetic resonance (NMR) play a leading role in yielding informative data for the identification of both known and unknown organic chemical compounds. Both have benefits and drawbacks that characterize their complementary usage in terms of sensitivity, reproducibility and structural information they are able to

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

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

† Electronic Supplementary Information (ESI) available: The file nmrfilter.zip contains all code described in Section 2.2. The data in realspectrum.csv and compounds.smi are for P. boldus. The spectral data of P. boldus and the caffeine/ferulic acid mixture are provided as well. See DOI: 10.1039/b000000x/

‡ Corresponding author: stefan.kuhn@dmu.ac.uk.

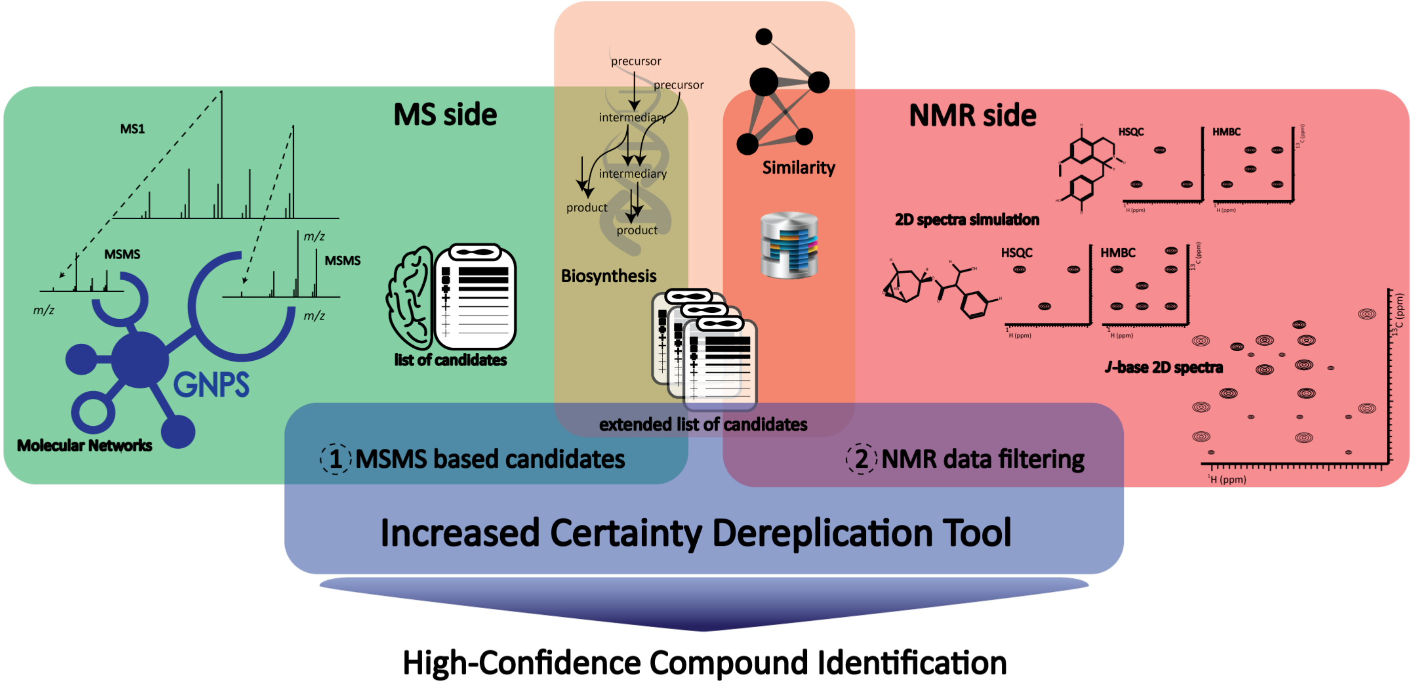


Fig. 1 Overall approach from the MSMS analysis and compilation of the possible candidates to the NMR simulations and matching.

provide. Whereas MS shows high sensitivity and accuracy, but low reproducibility, NMR shows low sensitivity, high reproducibility and efficiency to unambiguously elucidate complex structures. If a single compound is analysed, the spectrum can be interpreted and will reveal its structure. A molecular structure can be inferred from the peaks in the so-called spectrum acquired from both MS and NMR. In the case of mixtures, the spectrum corresponds to the spectra of all compounds in one analysis. So a direct interpretation as the result of a molecule is not trivial since all signals from each component of the mixture will be shown.

1.1 Background

Mixtures analysis is a hot topic today within NP and metabolomics including modern and complex algorithms. Specifically for NP, the analysis of complex mixtures is often referred as dereplication due to its goal to quickly identify known compounds and prevent replicated results. Dereplication in NP was extensively reviewed elsewhere.[2] Open-access tools such as MZMine and[3] OpenChrom[4] enable complex processing of MS data and database matching using open-access or even user-defined databases for dereplication; closed source options from different companies are available as well but under copyright protection. Global Natural Products Social Molecular Networking (GNPS) is an important tool that calculates similarities networks among the fragmentograms and enables a crowdsourcing approach for dereplication.[5] It uses open-access databases for spectra matching and allows the user to submit data from putatively identified compounds to a local database. New workflows are under development for the use of *in silico* fragmentation for compound identification.[6] GNPS now includes a workflow for the use of *in silico* fragmentation of candidate structures, namely Network Annotation Propagation.[7] Regarding NMR compound identification, COLMAR[8] is a broadly used tool for metabolomics mainly focused on primary metabolites. It uses HMDB[1] and BMRB[9] as database and it offers an interactive web interface. The underlying technique (called DemixC) for NMR analysis uses full high-resolution TOCSY after covariance NMR to deconvolute pure spectra from redundant connectivity information from the cross peaks. Statistical techniques are then used to find correlated changes in the cross peaks and allows separation of the spectra of the individual compounds from the measured spectra.[10] The use of ^{13}C NMR for compound identification is a trend in the last decade[11–14], probably due to the increasing sensitivity of dedicated (micro- and nano-)probes. Undeniably the ^{13}C resonances are less affected by external parameters such as solvent, pH or temperature than ^{1}H resonances, but the low sensitivity of direct ^{13}C detection is still prohibitive. The INETA package was designed to use INADEQUATE NMR data using mainly BMRB[9] and assigned ^{13}C resonances as database for compound identification, but yet it is only feasible for ^{13}C labeled samples. Another interesting approach was developed for a computer-aided ^{13}C profile of NP that uses 1D ^{13}C NMR data and an *in-house* search algorithm based on simulated NMR data from predefined candidates.[15] Later, the same group extended this approach by using HMBC NMR data for compound identification using a community detection algorithm.[16] Differential analysis of 2D NMR spectra (DANS) compares spectra from different biological states. If some signals vary significantly between two spectra, they are assumed to come from compounds unique to a particular sample.[17]

	Tolerance	
RBER Resolution	^{13}C: 0, ^{1}H: 0	^{13}C: 0.2, ^{1}H: 0.02
0.2	3/7, 2/6, 1/5, 1/2, 15/1	1/28, 1/15, 5/2, 2/1
0.5	3/7, 2/6, 1/5, 1/2, 15/1	1/28, 1/10, 1/5, 5/2, 2/1
1	3/7, 2/6, 1/5, 1/2, 15/1	1/15, 1/13, 1/7, 1/5, 1/3, 5/2, 2/1
10	3/5, 2/4, 3/3, 3/2, 17/1	1/6, 2/5, 3/4, 2/3, 8/2, 5/1

Table 1 The clustering achieved for the HMBC cross peaks of a mixture of caffeine and ferulic acid. Various settings for the chemical shift tolerance and the resolution parameter of the RBER algorithm were tested. 1/28, 5/2... means there was 1 cluster with 28 elements, 5 with 2 elements etc.

2 Methods

2.1 Overview of the method

The overall method we present here starts with a general compound identification scheme using LC-ESI-MSMS and molecular networks as described elsewhere.[18] We also searched a list of expected compounds in the literature assuming chemotaxonomical relations within the plant species and genus as well as compounds from related biosynthetic pathways. Once we have identified a set of lead compounds using MS data and chemotaxonomic review, we searched for similar known compounds through Pubmed and list them together as possible candidates (Fig. 1). Thus, this approach filters compound by structural variations rather than by the molecular mass only. For the candidates identified, we predict HMBC and HSQC NMR spectra and compare them to the measured spectrum of the mixture for every candidate (see Section 2.3 for details). The goal is to design a redundant process to confirm the MS compound identification using 2D NMR and increase confidence. For this, we have designed output parameters, which calculate how well the simulated spectra fit to the measured data, and the candidates are ranked accordingly.

2.2 NMR Network Analysis

An NMR network analysis was suggested in[16]. This is based on the idea that in an HMBC spectrum cross peaks originating from one compound should either share the ^{13}C or the ^{1}H chemical shift. So an initial network is built from the long-range proton-carbon couplings. If a complete separation is not possible by this, a community clustering algorithm should separate the subnetworks for the individual compounds in the overall network, assuming that there should be more connections between cross peaks within one compound than to the other compounds. In[16] the RBER (Erdös-Rényi null-model) method is used, with the resolution parameter set to 0.2. This method divides the network into clusters based on the density of the connections inside the clusters compared to the density of connections to other clusters. It optimizes the clusters to have many connections within, but few to other clusters. Cross peaks from the same compound should share many chemical shifts. In contrast, different compounds should share chemical shifts rarely, by a combination of similar substructures and not enough resolution in the measurement. Therefore, the clusters should mostly correspond to cross peaks from one structure, not from several structures.

It is reported in[16] that this process yields as many clusters as there are compounds in the mixture, with each cluster corresponding to the cross peaks from one compound. The cross peaks from each cluster can then be matched against a database of compounds to identify the components of the mixture.

In order to verify the approach we tested it with the mixture of caffeine and ferulic acid described in Section 4. The HMBC spectrum yields 55 cross peaks. We firstly built a network of HMBC cross peaks using the 0.2/0.02 ppm tolerance and then applied the RBER using the suggested 0.2 resolution parameter. We then varied these parameters. Table 1 shows the numbers of clusters and their size derived with different resolution parameters for the RBER algorithm. As expected, we obtained two large clusters by setting the resolute parameter to 0.2. We listed the observed cross peaks (numbered from 1 to 55) in the clusters, and underline them if there is a matching cross peak for caffeine and overline them if there is a matching cross peak for ferulic acid. 42 of the 47 predicted chemical shifts are matched to a measured chemical shift. We get the following list:

$[\overline{1},\overline{2},\overline{3},\overline{4},\overline{6},\overline{7},\overline{8},\overline{9},\overline{10},\overline{11},\overline{12},\overline{13},\overline{14},\overline{15},\overline{16},\overline{17},\overline{18},\overline{20},21,\overline{22},\overline{23},\overline{38},\overline{39},\underline{45},$
$\overline{46},\overline{47},50,52]$

$[\overline{5},30,\underline{31},\underline{32},\underline{35},\underline{37},\underline{41},\underline{\overline{42}},44,\underline{\overline{48}},\underline{49},\underline{51},53,\overline{54},\overline{55}]$

$[19,40]$

$[24,25]$

$[26,27]$

$[\underline{28},29]$

$[33,34]$

$[\underline{36}]$

$[\underline{43}]$

We can see that the two large clusters correspond roughly to the two compounds, but the separation is not perfect. Furthermore, when using different parameters for resolution and tolerance, the number of clusters varies. We list the cross peaks for resolution 0.2 and tolerance 0 ppm for both axes in the same fashion as before, with the cross peaks matched marked by underline/overline, and visualize this in Fig. 2 (15 clusters with one cross peak have been left out for clarity):

$[\overline{5},\underline{31},\underline{32},\underline{35},\underline{36},\underline{37},\underline{41}]$

$[\overline{1},\overline{10},\overline{11},\overline{15},\overline{16},\overline{23},\overline{39}]$

$[\overline{4},\overline{7},\overline{8},\overline{9},\overline{13},\overline{14},\overline{17}]$

$[\overline{3},\overline{6},\overline{9},\overline{12},\overline{20},21]$

$[\underline{\overline{48}},\underline{49},\underline{51},53,\overline{54},\overline{55}]$

$[\overline{22},\overline{38},44,\overline{46},\overline{47}]$

$[24,25]$

There no clean separation, but we can still see a clustering pattern: In the clusters, most of the cross peaks belong to one compound.

Overall it is clear that the separation depends on the parameters. Furthermore, with more complex mixtures the best parameter setting may change. Finally, we found that the data processing and peak picking can influence the separation. On the other hand, even if a complete separation is not possible, the cross peaks for the compounds still fall mostly within specific clusters. They are not spread out over all clusters. We have also confirmed that some test compounds other than caffeine and ferulic acid do not show such a clustering pattern when tested against the caffeine and ferulic acid mixture. The cross peaks of those other compounds either do not match any cross peaks or spread out over all clusters.

2.3 NMR Filtering by Network Analysis

Considering our results when using the NMR network analysis, we believe that a reliable separation of compounds is not always possible. If this would be possible, each cluster could be matched against the predicted spectra. Even though a reliable separation is not possible, the clusters still somehow relate to the compounds. Since we have a list of possible candidates from the MS experiments, we have therefore devised a modified algorithm. This does not assume that one cluster represents one compound, but that clusters contain cross peaks belonging to one compound. Even if it would be possible to achieve full separation by fine-tuning the measurement and the data processing and peak picking, our results indicate that the full separation is quite sensitive and is not guaranteed to work. Therefore our method is designed to be more robust and less dependant on the quality of the data. We call our method NMR filtering by network analysis. If the clustering separates the compounds exactly, the simulated spectra should cover exactly one clusters. So the procedure in[16] is actually a special case of our algorithm.

Once a list of candidates has been generated using the MS analysis, the results are ranked according to the likelihood of their occurrence in the

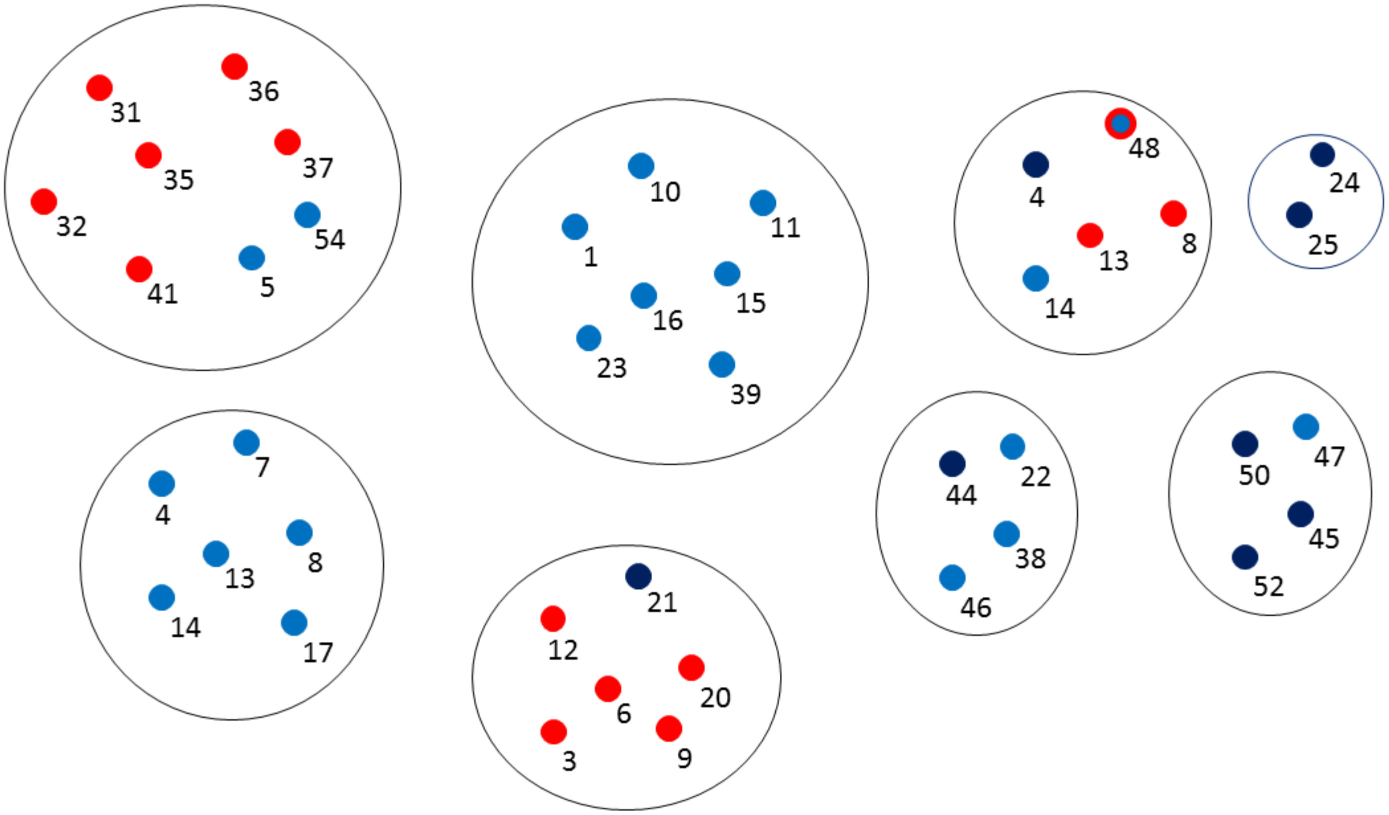


Fig. 2 The clusters derived from the HMBC spectrum of caffeine and ferulic acid with tolerances set to 0 for both chemical shifts and the RBER resolution set to 0.2, with the compounds mapped onto them. Cross eaks for ferulic acid are blue, those for caffeine red, cross peaks mapped to both are in both colours. Cross peaks used for none are in black.

2D NMR spectra. Core ideas presented in[16] are used for this, but they are extended and generalized in the present approach that uses established techniques, but modifies them by introducing new elements. The core steps of our approach are:

- We use HSQC and HMBC spectra, measured and peak-picked as explained in Section 4. The cross peaks of both spectra were put into a single list, with the ^{13}C chemical shift being the first dimension and the ^{1}H chemical shift being the second dimension.
- Every cross peak is a node in the NMR network we built in the next step. An edge between two nodes is added to the network if two cross peaks share a chemical shift on the ^{13}C or ^{1}H axis. A tolerance of 0.2 ppm for ^{13}C chemical shifts and 0.02 ppm for ^{1}H chemical shifts is applied here. These values have been found experimentally and can be changed if desired. As cross peaks from the same compound should share either the ^{13}C or the ^{1}H chemical shift value(s) with other cross peaks from the same compound, this gives an initial network.
- The resulting network is analysed using the RBER algorithm. The resolution parameter is set to 0.2. Again, this can be changed. As explained, inside the clusters produced, the cross peaks should originate predominantly from one compound, even if a complete separation is not possible.
- We then predict the HSQC and HMBC spectra for the candidate structures derived from MSMS. The combined spectra for each of the compounds are then mapped onto the measured spectrum. From the mapping, we calculate two measures: a) the distance of the simulated spectrum to the measured spectrum and b) the distribution of the cross peaks matched in the measured spectrum within the clusters calculated in the previous steps. For details of the calculation see the description of the implemenation below.
- We normalize both measures to range from 0 to 1 and use the average of the two measures as the likelihood of a compound to be part of the mixture.

In this algorithm, we map the cross peaks of the simulated spectrum of each candidate onto the whole spectrum and calculate the distribution over all clusters. Ideally, the cross peaks should cover some clusters completely and not have any cross peaks in the remaining clusters. So we have the distribution in the clusters and the distance of the simulated spectrum to the best match in the measured spectrum for each candidate as indication of how likely the candidate is to occur in the mixture. In order to improve the clustering, we include HSQC and HMBC spectra in our clustering (as opposed to[16], which uses HMBC only). All of these have ^{13}C-^{1}H cross peaks, which are treated the same, forming one network, to which the clustering is applied. The spectrum simulation is also done for HMBC and HSQC spectra and these cross peaks are mapped onto the combined spectra.

Our approach is illustrated in Section 2.2 and Figure 2. They demonstrate that the cross peaks for the compounds fall mostly within specific clusters. This is true even if a complete separation is not achieved.

We have implemented the described procedure as a set of Python script, including a Java program to do the prediction, and a shell script to run the overall procedure. Data are transferred between scripts via text files. This is primarily intended as proof of concept, a full application is part of the future work. The detailed algorithm for the NMR ranking is as follows:

- For each candidate structure originating from the MSMS step we simulate the combined HSQC and HMBC spectra using the prediction mechanism of nmrshiftdb2[19]. The Java code in `simulate.jar` extracts pairs of atoms from the molecule, which are assumed to generate a cross peak in one of the spectra, and writes the pair of chemical shifts of these atoms into a peaklist. For HSQC, cross peaks are built for all atoms pairs one bond away, and for HMBC, for all atom pairs two or three bonds away. Couplings and intensities are currently not included, the cross peaks are based on topology only. Experience shows that this gives a sufficient approximation. The chemical shift prediction is based on HOSE codes, uses solvents when possible, and respects wedge bonds if data are available.[20]

- We form a single list of cross peaks out of the HSQC and HMBC spectra measured for the mixture. This list is provided to the `clustering.py` script, which builds a network as described, using the tolerances from the `nmrproc.properties` file.
- The network generated is processed by `clustserlouvain.py`. This applies the RBER algorithm, using the louvain library.[21] The resolution value is taken from the `nmrproc.properties` file. The result is a list of clusters, containing all cross peaks from the measured spectra in some cluster.
- For every simulated spectrum, we find the nearest matching cross peaks in the measured spectrum. This is done by calculating the distance between each cross peak in the simulated spectrum and each cross peak in the measured spectrum. The formula for this is as follows:

$$distance(peak_1, peak_2) = (abs(peak_{1_x} - peak_{2_x}) + abs(peak_{1_y} - peak_{2_y}) * 10)^2 \quad (1)$$

  This squares the distance between the two cross peaks on the ^{1}H and ^{13}C axis and adds them. The ^{1}H chemical shift is multiplied by 10 to normalize the range, assuming ^{13}C ranges from 0 to 200 ppm and ^{1}H from 0 to 20 ppm. The factor 10 is commonly used, e. g. in[16] the tolerance for carbon chemical shifts is 1.5 ppm and for hydrogen chemical shifts it is 0.15 ppm. The squaring includes variance and bias, similar to the mean squared error in statistics. This gives us a matrix of size $n * m$, where n is the number of cross peaks in the measured spectrum and m the number of cross peaks in the simulated spectrum. We then use the function `linear_sum_assignment` from the `scipy.optimize` package to find the minimal combination of these costs, which assigns exactly one cross peak to every cross peak in the measured spectrum. The sum of the costs of this minimal combination is the distance of the simulated spectrum to the measured spectrum, which is our first reliability measure. As opposed to other methods[22], we do not have a fixed limit for cross peaks to match, rather we search for a best match and calculate the distance. Together with the squared distance, this should give us a robust mapping.
- We have previously created the clusters containing the measured cross peaks. In the previous step, we have mapped each simulated cross peak onto one measured cross peaks. Therefore, we can now calculate the fraction of cross peaks in each clusters, onto which a simulated cross peak is mapped. Our distance measure will map each simulated cross peak onto some measured peak, even if they are very much apart. For the distance measure this is not a problem, since it will mean a very high distance value in cases of bad matches, which in turn means the compound will not be considered a good match. For the clustering step, we only use mappings where the distance is less than 9 which corresponds to a value of 1.5 ppm for the ^{13}C chemical shift and 0.15 ppm (remember the factor of 10) for the ^{1}H chemical shift, which are the cut offs used in[16]. This step gives us n decimal numbers between 0 and 1 (since it represents the fraction of mapped peaks, which is between 0 for no mapping and 1 for all peaks mapped), n being the number of clusters. We then calculate the standard deviation, using the `std` method of the `numpy` package of these numbers for each simulated spectrum. The standard deviation is the second reliability measure.
- In the last step, both reliability measures are normalized to range from 0 (best) to 1 (worst). For each simulated spectrum, and therefore for each candidate compound, they are added and the compounds are ranked by this combined reliability measure, the compound with the lowest value being the most likely candidate.

It should be noted that the current code is not optimized for performance. Running it on a laptop with an Intel Core i5 6300U CPU for the P. boldus mixture discussed in the next section takes around 4 minutes. This will be improved in the planned application, but for this type of task a very quick solution cannot be expected, given the amount of information involved.

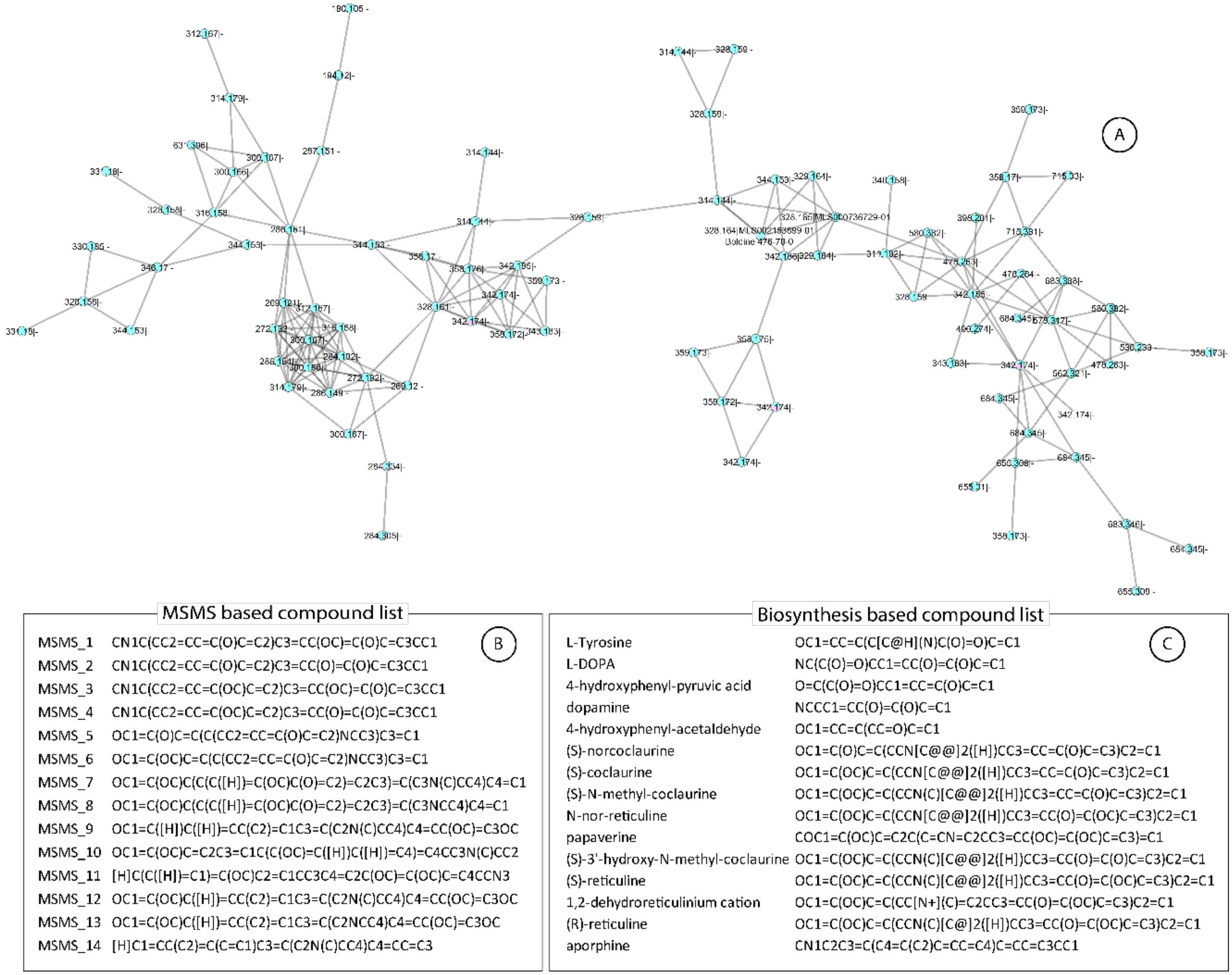
3 Results

We collected MS data in high resolution under ddMS2 Top3 experiments to yield close to 5000 scan in more than 2500 Mb file. We converted the raw data to .mzXML for network calculations using the GNPS web system and, then, we used Cytoscape for visualization and further analysis (Fig. 3, section A). From the MS spectra, we could visualize high intensity key features that would indicate well-known components of the alkaloidic fraction of *P. boldus*. Boldine (at *m/z* 328.15), coclaurine (at *m/z* 286.14), and norreticuline (at *m/z* 316.15) are well-known components of the aporphine-like pool of alkaloids from this species[23]; other close related (delta-*m/z* 14, 12 and 16) features (at *m/z* 300.15, 342.17 and 358.16) are also displayed at the MS scan (Supplementary Information). Nonetheless, the use of molecular networking for structure elucidation enable the enrichment of the list of candidates, and adding others that are close related or of expected occurrence. The GNPS processed data can be accessed here.* GNPS promptly identified boldine within a network of 94 nodes; this finding enable us to relay the information and elucidate the possible structure for the close related nodes. 14 structures were suggested in this stage (Fig. 3, section B). Thus, we identified a core structure as being of an aporphine-like alkaloid and used that to extend the list of candidates with similar known compounds (from Pubmed) and other that plays a role in their most accepted biosynthesis pathway (Fig. 3, section C).[24] We took the compounds from the MS side of the method and listed them as SMILES structures as preparations for the NMR filtering. Note that the sensitivity gap between NMR and MS is anticipated and will be reflected naturally by the eventual annotation of fewer compounds after NMR analysis side of the method. This is due to the peak intensity of MS data which is structure-dependent and vary highly according to the ionization technique. In contrast, peak intensity in NMR is mainly dependent on the spin concentration, ^{1}H in the case discussed here.

The HMBC and HSQC spectra for *P. boldus* gave 1034 cross peaks. Running the clustering algorithm on these yields 193 clusters. The largest has 257 cross peaks, 75 clusters have one peak, the other clusters are somewhere in between in size. The average size of the clusters is 4.36, the median is 2. We simulate the spectra for the compounds derived from the MS step. Calculating the similarity and clustering for them, gives a ranking for the compounds. The first ten compounds and the last compound are as follows:

```
1: OC1=C(OC)C=C(C(CC2=CC=C(O)C=C2)NCC3)C3=C1,
     distance: 0.21, standard deviation: 1.00; 1-[(4-
     hydroxyphenyl)methyl]-7-methoxy-1,2,3,4-
     tetrahydroisoquinolin-6-ol
2: OC1=C(O)C=C(CCN[C@@]2([H])CC3=CC=C(O)C=C3) C2=C1,
     distance: 0.23, standard deviation: 0.99; (S)-
     Norcoclaurine
3: OC1=C(O)C=C(C(CC2=CC=C(O)C=C2)NCC3)C3=C1, distance
     : 0.24, standard deviation: 0.98; Norcoclaurine
4: CC1CN(C)C2CC3=CC=CC=C3C4=C2C1=CC=C4, distance:
     0.26, standard deviation: 0.99; 4,6-dimethyl
     -5,6,6a,7-tetrahydro-4H-dibenzo[de,g]quinoline
5: OC1=C(OC)C=C(CCN[C@@]2([H])CC3=CC(O)=C(OC)C =C3)C2
     =C1, distance: 0.20, standard deviation: 0.84; (S
     )-Norreticuline
```

* See the link `https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=4275dd938bdf4eea8f30a59afdcfc671`



MSMS based compound list

MSMS_1	CN1C(CC2=CC=C(O)C=C2)C3=CC(OC)=C(O)C=C3CC1
MSMS_2	CN1C(CC2=CC=C(O)C=C2)C3=CC(O)=C(O)C=C3CC1
MSMS_3	CN1C(CC2=CC=C(OC)C=C2)C3=CC(OC)=C(O)C=C3CC1
MSMS_4	CN1C(CC2=CC=C(OC)C=C2)C3=CC(O)=C(O)C=C3CC1
MSMS_5	OC1=C(O)C=C(C(CC2=CC=C(O)C=C2)NCC3)C3=C1
MSMS_6	OC1=C(OC)C=C(C(CC2=CC=C(O)C=C2)NCC3)C3=C1
MSMS_7	OC1=C(OC)C(C(C([H])=C(OC)C(O)=C2)=C2C3)=C(C3N(C)CC4)C4=C1
MSMS_8	OC1=C(OC)C(C(C([H])=C(OC)C(O)=C2)=C2C3)=C(C3NCC4)C4=C1
MSMS_9	OC1=C([H])C([H])=CC(C2)=C1C3=C(C2N(C)CC4)C4=CC(OC)=C3OC
MSMS_10	OC1=C(OC)C=C2C3=C1C(C(OC)=C([H])C([H])=C4)=C4CC3N(C)CC2
MSMS_11	[H]C(C([H])=C1)=C(OC)C2=C1CC3C4=C2C(OC)=C(OC)C=C4CCN3
MSMS_12	OC1=C(OC)C([H])=CC(C2)=C1C3=C(C2N(C)CC4)C4=CC(OC)=C3OC
MSMS_13	OC1=C(OC)C([H])=CC(C2)=C1C3=C(C2NCC4)C4=CC(OC)=C3OC
MSMS_14	[H]C1=CC(C2)=C(C=C1)C3=C(C2N(C)CC4)C4=CC=C3

Biosynthesis based compound list

L-Tyrosine	OC1=CC=C(C[C@H](N)C(O)=O)C=C1
L-DOPA	NC(C(O)=O)CC1=CC(O)=C(O)C=C1
4-hydroxyphenyl-pyruvic acid	O=C(C(O)=O)CC1=CC=C(O)C=C1
dopamine	NCCC1=CC(O)=C(O)C=C1
4-hydroxyphenyl-acetaldehyde	OC1=CC=C(CC=O)C=C1
(S)-norcoclaurine	OC1=C(O)C=C(CCN[C@@]2([H])CC3=CC=C(O)C=C3)C2=C1
(S)-coclaurine	OC1=C(OC)C=C(CCN[C@@]2([H])CC3=CC=C(O)C=C3)C2=C1
(S)-N-methyl-coclaurine	OC1=C(OC)C=C(CCN(C)[C@@]2([H])CC3=CC=C(O)C=C3)C2=C1
N-nor-reticuline	OC1=C(OC)C=C(CCN[C@@]2([H])CC3=CC(O)=C(OC)C=C3)C2=C1
papaverine	COC1=C(OC)C=C2C(C=CN=C2CC3=CC(OC)=C(OC)C=C3)=C1
(S)-3'-hydroxy-N-methyl-coclaurine	OC1=C(OC)C=C(CCN(C)[C@@]2([H])CC3=CC(O)=C(O)C=C3)C2=C1
(S)-reticuline	OC1=C(OC)C=C(CCN(C)[C@@]2([H])CC3=CC(O)=C(OC)C=C3)C2=C1
1,2-dehydroreticulinium cation	OC1=C(OC)C=C(CC[N+](C)=C2CC3=CC(O)=C(OC)C=C3)C2=C1
(R)-reticuline	OC1=C(OC)C=C(CCN(C)[C@]2([H])CC3=CC(O)=C(OC)C=C3)C2=C1
aporphine	CN1C2C3=C(C4=C(C2)C=CC=C4)C=CC=C3CC1

Fig. 3 Main molecular network and list of candidates.

```
6: OC1=C(OC)C(C(C([H])=C(OC)C(O)=C2)=C2C3)=C( C3N(C)
    CC4)C4=C1, distance: 0.13, standard deviation:
    0.73; (+)−(S)−Boldine
7: OC1=C(OC)C=C(CCN[C@@]2([H])CC3=CC=C(O)C=C3)C2=C1,
    distance: 0.23, standard deviation: 0.82;
    Coclaurine
8: CCC1CN(C)C2CC3=CC=CC=C3C4=C2C1=CC=C4, distance:
    0.26, standard deviation: 0.81; 4−ethyl−6−methyl
    −5,6,6a,7−tetrahydro−4H−dibenzo[de,g]
9: OC1=C(OC)C=C(CCN(C)[C@@]2([H])CC3=CC(O)=C(O) C=C3)
    C2=C1, distance: 0.16, standard deviation: 0.67;
    (S)−3−Hydroxy−N−methylcoclaurine
10: NCCC1=CC(O)=C(O)C=C1, distance: 0.00, standard
    deviation: 0.47; Dopamine
...
66: O=C(C(O)=O)CC1=CC=C(O)C=C1, distance: 0.99,
    standard deviation: 0.00; 4−Hydroxyphenylpyruvate
```

The first hits have high standard deviations (due to normalizing 1 is maximum), meaning they fall into a low number of clusters. They also have relatively high distance to the measured spectrum (0 being optimal). Lower in the list, we get better similarities, but still relatively high clustering. The last hit has high distance and a low standard deviation, making it a highly unlikely candidate. This ranking validates the MS-only based dereplication yielding higher confidence to the result. The successful identification of such aporphine alkaloids (1-[(4-hydroxyphenyl)methyl]-7-methoxy-1,2,3,4-tetrahydroisoquinolin-6-ol, norcoclaurine, 4,6-dimethyl-5,6,6a,7-tetrahydro-4H-dibenzo[de,g]quinoline, norreticuline, boldine, coclaurine, 4-ethyl-6-methyl-5,6,6a,7-tetrahydro-4H-dibenzo[de,g] and 3-Hydroxy-N-methylcoclaurine) using both MS and NMR matches the experimental results for *P. boldus* from the previous studies[23].

4 Experimental section

4.1 Chemicals

HPLC grade Methanol and Ethyl Acetate, LC-MS grade formic acid and HCl and NaOH were acquired from Tedia-Brazil (Rio de Janeiro, RJ, Brazil); D_2O (99.0%), methanol-d^4 and chloroform-d^1 were acquired from Cambridge Isotope laboratory, Inc. (Andover, MA, USA); caffeine and ferulic acid were acquired from Sigma-Aldrich (St. Louis, MO, United States). Deionized water was purified by a Millipore Milli-Q Gradient A 10 System (Burlington, MA, USA).

4.2 Plant Material and Sample preparation

Peumus boldus dry leaves from different brands were purchased from different commercial locations in Rio de Janeiro (RJ, Brazil). Samples of 1 g of each were combined and an aliquot (1 g) was saved for the extraction. This representative aliquot was extracted with aqueous HCl 0.02M (15 ml) at pH 2.5 under ultrasound for 5 minutes. Then, three successively liquid-liquid extractions were performed with 5 ml of Ethyl Acetate. The pH of the aqueous phase was increased to 9 with 1 ml of aqueous NaOH (1 M) and three successive extractions were again performed with 5 ml of Ethyl Acetate. The combined organic phases was concentrated to dryness under vacuum to yield 22.6 mg of a crude alkaloid extract. This preparation was made in three replicates. The final samples were divided into 2 aliquots each. 20% of it was resuspended in methanol for LC-MSMS analysis and the remaining 80%, in chloroform-d^1 for the NMR analysis; for the NMR analysis the replicates were combined. The caffeine and ferulic acid mixture was prepared in methanol-d^4 as concentrated samples, centrifuged and transferred to 3 mm NMR tubes.

4.3 Liquid Chromatography-Tandem Mass Spectrometry Analysis

Ultra-high performance liquid chromatography analysis was performed on a 1260 Infinity Liquid Chromatography system (Agilent) consisting of a quaternary solvent delivery pump and a column oven compartment. Samples (10 μL) were injected using and separated on an Agilent Extend-C18 column (2.1x50 mm, 1.8 μm particle size) at 300 μL min^{-1} maintained at 40 ^{o}C. The mobile phase consisted of (A) 0.1% formic acid and (B) 0.1% formic acid in methanol in gradient elution mode (0 min 15% B; 0-10 min 100% B; 10-18 min, 15% B; 18.5-25 min 15%). The UHPLC system was coupled to a Q-TOF high resolution and accurate mass spectrometer (Agilent) equipped with an electrospray ion source (Dual ESI; Agilent) operating in positive ionization mode. Source ionization parameters were: spray voltage 3.5 kV; capillary temperature 350 ^{o}C; gas flow 10 l/min; nebulizer 25 psi; skimmer1 65; isolation width MS/MS medium (˜ 4 amu); fixed collision energy for MS/MS 30. Samples were analysed in the scan range of m/z 100 to 1700 (for MS and MS/MS) at a scan rate of 3 spectra/sec followed by data-dependent MSMS (ddMS2 Top3 experiments) at a scan rate of 2 spectra/sec. The acquired data were converted to mzML or mzXML files using the software MSConvert (ProteoWizard; proteowizard.sourceforge.net/tools.shtml). GNPS network parameters were: MS Fragment Ion Tolerance: 0.02; MS/MS Fragment Ion Tolerance: 0.02; Minimum MS/MS Peak Intensity: 0.0; Run MSCluster: on; Minimum Consensus Cluster Size: 1; Minimum Matched Peaks in Network Edge: 4; Minimum MS/MS cosine score in Network Edge: 0.65; Number of Neighbors to Retain in Network: 10; Maximum Connected Component Size: 100. The resulted networks were plotted using the Cytoscape software (http://www.cytoscape.org).

4.4 Nuclear Magnetic Resonance Analysis

NMR data was collected using a 800 MHz and a 600 MHzBruker Avance III equipped with a 1.7 mm TCI cryoprobe and a 5 mm DCH D/H-C carbon cryoprobe, respectively. The pulse sequence hsqcedetgpsp.3 under non-uniform sampling mode (35% of NUS amount and 896 NUS points; 4096 and 5120 points for F2 and F1, respectively; 28.45 points/ppm) was used to acquire the edited HSQC data (24 scans, optimized for $^1J_{CH}$ =145 Hz; 18h 15 min), and hmbcetgpl3nd under non-uniform sampling mode (30% of NUS amount and 768 NUS points; 4096 and 5120 points for F2 and F1, respectively; 23.27 points/ppm) for the HMBC data (24 scans, optimized for$^3J_{CH}$=8 Hz; 14h 22 min). For the test sample (caffeine plus ferulic acid) the NMR data was collected using hsqcedetgpsp.3 under non-uniform sampling mode (1024 and 256 points for F2 and F1, respectivelly) was used to acquire the edited HSQC data (4 scans, optimized for $^1J_{CH}$ =145 Hz; 8 min), and hmbcetgpl3nd under non-uniform sampling mode (4096 and 256 points for F2 and F1, respectivelly) for the HMBC data (4 scans, optimized for $^3J_{CH}$=8 Hz; 7 min).

5 Conclusion

We have demonstrated a method to infer a list of candidate compounds for complex natural product mixtures. Our method combines MS and NMR techniques to give confidence in the results. The MS step yields a relatively broad result, ensuring coverage of all possible compounds. The NMR step does not rely on predefined libraries, but ranks the suggestions by using their predicted NMR spectra. The prediction can be done for a range of naturally occurring products with a reasonable average error.[20] We found that a full distinction of the compounds in the spectrum is not needed to rank the candidates. Since a full distinction is difficult and in many cases not possible, we consider the combination of prefiltering and ranking a promising approach. It gives reasonable results even if the peak data are not optimal, due to problems in measurement or data processing. There are indications that the results provide a good match with the actual compounds, but more work to verify this is needed. In particular, larger datasets will be examined by the authors.

We have used an artificial mixture to demonstrate the NMR filtering step and have demonstrated the overall approach using an alkaloid enriched extract of *P. boldus*. A major advantage is that no special sample preparation or experiments are needed. Both the MS and NMR experiments are standard and can be used almost as default. Even though better resolution and higher sensitivity will improve the results, the use of older or less sophisticated equipment is still possible. Furthermore, once the experiments are performed, the processing is relatively quick and will be even more with more automation, which we intend to make possible.

5.1 Future work

The process as presented in this paper is only partially automated. The computational parts are currently done in a command line interface without possibility for user interaction. We aim to increase automation and make the interface more user-friendly in a next step. We consider integrating the program either into a workflow tool like KNIME or a platform like Bioclipse. The inclusion of $^nJ_{HH}$ data in the NMR network analysis will be part of this.

Concomitantly, we are applying this approach to other samples and fractions for a broader range of applications; mainly the NMR filter will benefit significantly by reducing the complexity and the dynamic range with a low-resolution fractionation step. The goal is to establish a source independent tool for dereplication of NP to be used as driving force towards novelty discovery. The scripts will be made freely available and it will enable data submission to databases as integral part. New samples of terrestrial plant, marine organisms, microorganisms, fungi and corals are some of the examples to be exploited in the near future.

6 Conflicts of interest

There are no conflicts of interest to declare.

7 Acknowledgments

We thanks National Council for the Improvement of Higher Education (CAPES, Brazil) for the financial support. RMB would like to thanks Dr Arthur Edison for all the support in this research topic and for NMR time at the Complex Carbohydrate Research Center (UGA, USA).

References

1 D. S. Wishart, Y. D. Feunang, A. Marcu, A. C. Guo, K. Liang, R. Vazquez-Fresno, T. Sajed, D. Johnson, C. Li, N. Karu, Z. Sayeeda, E. Lo, N. Assempour, M. Berjanskii, S. Singhal, D. Arndt, Y. Liang, H. Badran, J. Grant, A. Serra-Cayuela, Y. Liu, R. Mandal, V. Neveu, A. Pon, C. Knox, M. Wilson, C. Manach and A. Scalbert, *Nucleic Acids Res.*, 2018, **46**, D608–D617.

2 J. Hubert, J.-M. Nuzillard and J.-H. Renault, *Phytochemistry Reviews*, 2017, **16**, 55–95.

3 F. Olivon, G. Grelier, F. Roussi, M. Litaudon and D. Touboul, *Analytical Chemistry*, 2017, **89**, 7836–7840.

4 P. Wenig and J. Odermatt, *BMC Bioinformatics*, 2010, **11**, 405.

5 M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. V. Melnik, M. J. Meehan, W. T. Liu, M. Crusemann, P. D. Boudreau, E. Esquenazi, M. Sandoval-Calderon, R. D. Kersten, L. A. Pace, R. A. Quinn, K. R. Duncan, C. C. Hsu, D. J. Floros, R. G. Gavilan, K. Kleigrewe, T. Northen, R. J. Dutton, D. Parrot, E. E. Carlson, B. Aigle, C. F. Michelsen, L. Jelsbak, C. Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B. T. Murphy, L. Gerwick, C. C. Liaw, Y. L. Yang, H. U. Humpf, M. Maansson, R. A. Keyzers, A. C. Sims, A. R. Johnson, A. M. Sidebottom, B. E. Sedio, A. Klitgaard, C. B. Larson, C. A. B. P, D. Torres-Mendoza, D. J. Gonzalez, D. B. Silva, L. M. Marques, D. P. Demarque, E. Pociute, E. C. O'Neill, E. Briand, E. J. N. Helfrich, E. A. Granatosky, E. Glukhov, F. Ryffel, H. Houson, H. Mohimani, J. J. Kharbush, Y. Zeng, J. A. Vorholt, K. L. Kurita, P. Charusanti, K. L. McPhail, K. F. Nielsen, L. Vuong, M. Elfeki, M. F. Traxler, N. Engene, N. Koyama, O. B. Vining, R. Baric, R. R. Silva, S. J. Mascuch, S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P. G. Williams, J. Dai, R. Neupane, J. Gurr, A. M. C. Rodriguez, A. Lamsa, C. Zhang, K. Dorrestein, B. M. Duggan, J. Almaliti, P. M. Allard, P. Phapale, L. F. Nothias, T. Alexandrov, M. Litaudon, J. L. Wolfender, J. E. Kyle, T. O. Metz, T. Peryea, D. T. Nguyen, D. VanLeer, P. Shinn, A. Jadhav, R. Muller, K. M. Waters, W. Shi, X. Liu, L. Zhang, R. Knight, P. R. Jensen, B. O. Palsson, K. Pogliano, R. G. Linington, M. Gutierrez, N. P. Lopes, W. H. Gerwick, B. S. Moore, P. C. Dorrestein and N. Bandeira, *Nat. Biotechnol.*, 2016, **34**, 828–837.

6 P.-M. Allard, T. Péresse, J. Bisson, K. Gindro, L. Marcourt, V. C. Pham, F. Roussi, M. Litaudon and J.-L. Wolfender, *Analytical Chemistry*, 2016, **88**, 3317–3323.

7 R. R. da Silva, M. Wang, L.-F. Nothias, J. J. J. van der Hooft, A. M. Caraballo-Rodríguez, E. Fox, M. J. Balunas, J. L. Klassen, N. P. Lopes and P. C. Dorrestein, *PLOS Computational Biology*, 2018, **14**, 1–26.

8 K. Bingol, L. Bruschweiler-Li, C. Yu, A. Somogyi, F. Zhang and R. Brüschweiler, *Analytical Chemistry*, 2015, **87**, 3864–3870.

9 J. F. Doreleijers, S. Mading, D. Maziuk, K. Sojourner, L. Yin, J. Zhu, J. L. Markley and E. L. Ulrich, *Journal of Biomolecular NMR*, 2003, **26**, 139–146.

10 F. Zhang and R. Brüschweiler, *ChemPhysChem*, **5**, 794–796.

11 A. Bruguière, S. Derbré, C. Coste, M. L. Bot, B. Siegler, S. T. Leong, S. N. Sulaiman, K. Awang and P. Richomme, *Fitoterapia*, 2018, **131**, 59 – 64.

12 A. Botana, P. W. Howe, V. Caër, G. A. Morris and M. Nilsson, *Journal of Magnetic Resonance*, 2011, **211**, 25 – 29.

13 A. Mäkelä, I. Kilpeläinen and S. Heikkinen, *Journal of Magnetic Resonance*, 2010, **204**, 124 – 130.

14 C. S. Clendinen, C. Pasquel, R. Ajredini and A. S. Edison, *Analytical Chemistry*, 2015, **87**, 5698–5706.

15 J. Hubert, J.-M. Nuzillard, S. Purson, M. Hamzaoui, N. Borie, R. Reynaud and J.-H. Renault, *Analytical Chemistry*, 2014, **86**, 2955–2962.

16 A. Bakiri, J. Hubert, R. Reynaud, C. Lambert, A. Martinez, J. H. Renault and J. M. Nuzillard, *J Chem Inf Model*, 2018, **58**, 262–270.

17 F. C. Schroeder, D. M. Gibson, A. C. Churchill, P. Sojikul, E. J. Wursthorn, S. B. Krasnoff and J. Clardy, *Angew. Chem. Int. Ed. Engl.*, 2007, **46**, 901–904.

18 L.-F. Nothias, M. Nothias-Esposito, R. da Silva, M. Wang, I. Protsyuk, Z. Zhang, A. Sarvepalli, P. Leyssen, D. Touboul, J. Costa, J. Paolini, T. Alexandrov, M. Litaudon and P. C. Dorrestein, *Journal of Natural Products*, 2018, **81**, 758–767.

19 S. Kuhn and N. E. Schlorer, *Magn Reson Chem*, 2015, **53**, 582–589.

20 S. Kuhn, B. Egert, S. Neumann and C. Steinbeck, *BMC Bioinformatics*, 2008, **9**, 400.

21 *louvain. PyPI*, `https://pypi.org/project/louvain/`.

22 K. Wolfram, A. Porzel and A. Hinneburg, Knowledge Discovery in Databases: PKDD 2006, Berlin, Heidelberg, 2006, pp. 650–658.

23 G. Fuentes-Barros, S. Castro-Saavedra, L. Liberona, W. Acevedo-Fuentes, C. Tirapegui, C. Mattar and B. K. Cassels, *Fitoterapia*, 2018, **127**, 179 – 185.

24 *Alkaloids*, John Wiley & Sons, Ltd, 2001, ch. 6, pp. 291–403.