Easy Structural Dereplication of Natural Products by Means of Predicted Carbon-13 Nuclear Magnetic Resonance Spectroscopy Data

Stefan Kuhn, Jean-marc Nuzillard

Easy structural dereplication of natural products by means of predicted carbon-13 nuclear magnetic resonance spectroscopy data

Stefan Kuhn,*[a] Jean-Marc Nuzillard*[b]

The present article reports the creation and usage of a general natural product database for the structural dereplication of natural products. This database, acd_lotusv7, derives from the LOTUS natural products database as the sole source of chemical structures. Database construction also relies on the commercial "ACD/C+H Predictors and DB" software for the prediction of the carbon-13 nuclear magnetic resonance (NMR) spectral data associated with structures. The linkage of each natural compound with a Wikidata resource identifier already present in LOTUS accelerates the access to the primary literature data such as biologic origin and bibliographic references. The open source nmrshiftdb2 web interface and search engine provide a simple and free way to retrieve compound structures stored in acd_lotusv7 from carbon-13 data and to analyze search results. Dereplication is illustrated by the easy and free retrieval of the structure of three natural compounds of low, medium, and high complexity from published lists of carbon-13 NMR chemical shifts.

Introduction

The concept of dereplication is invoked each time a way is found to avoid repeating a task that was already performed. In the field of the organic chemistry of natural products, a compound produced by different living organisms may have been purified many times by different research teams resulting in repeated structure elucidation works.[1] Purification dereplication avoids the repetitive search for an appropriate compound isolation method. Structural dereplication avoids the recording of detailed spectroscopic data and their interpretation for compounds already reported in the scientific literature.

Mixture analysis as reported by J. Hubert and collaborators combined purification and structural dereplication in a single workflow called CaraMel.[2] With this procedure, the isolation of all the constituent of a mixture to a high purity level is not required for compound identification. Structural dereplication is achieved within the CaraMel procedure by looking into a locally developed and enriched database for compounds that match with a list of experimental ^{13}C nuclear magnetic resonance (NMR) chemical shift values. The database contains structures that were accumulated through bibliographic searches driven by the taxonomic classification of the studied organisms. Each compound in this database is linked to a list of ^{13}C NMR predicted chemical shifts. The structural and spectroscopic data management is ensured by a software acquired from Advanced Chemistry Development, Inc. (hereafter, ACD/Labs, https://www.acdlabs.com/) called ACD/C+H Predictors and DB, hereafter referred to as the ACD/Labs software, even though the ACD/Labs company provides many other software products.

Supplementing a collection of small molecule structures with predicted NMR data by means of the ACD/Labs software is a tedious operation because there is by design no simple way to deal with batches of structures. Prediction quality is high but operations on large amounts of structures is unpractical. More precisely, prediction in ACD/Labs software may be carried out by two different procedures. The accurate and slow one relies on the search of molecular fragments within an internal database. Besides, a predictor is dedicated to the verification of experimental NMR chemical shifts values assigned by user to molecular structures. The latter predictor (hereafter called the validation algorithm) operates in an unsupervised way on collections of molecular structures for which a previously user–supplied NMR assignment exists. The output of the validation algorithm can be reused by a dedicated software in order to update the initially given atom to chemical shift assignments so that the predicted ones appear as if they were experimental. This way of supplementing a collection of structures with chemical shift values by means of the ACD/Labs validation algorithm was already reported in the context of the construction of small–sized, taxonomically focused databases.[3]

The present article reports the creation and usage of a general natural product database for structural dereplication. This database, acd_lotusv7, derives from the open–source LOTUS natural products database[4] and can be queried through the open–source web interface and search engine of nmrshiftdb2.[5] Natural product chemists can thus access published knowledge easierly than ever. The use of ^{13}C NMR spectroscopy as primary data source for compound identification and the advantages of resorting on predicted chemical shift values instead of experimental ones have already been discussed in previous works.[2,6] The overall number of identified natural products is not known with high accuracy as it increases every day but is evaluated as a few hundred thousands.[7] The LOTUS database is not presently comprehensive but is steadily enriched with new structures, making of it and of acd_lotus evolving tools for

[a] S. Kuhn*
Institute of Computer Science, University of Tartu. Narva mnt 18, 51009 Tartu, Estonia, formerly at School of Computer Science and Informatics, De Montfort University, Leicester, UK
E-mail: stefan.kuhn@ut.ee

[b] J.-M. Nuzillard*
CNRS, Université de Reims Champagne-Ardenne, ICMR, Reims, France
E-mail: jm.nuzillard@univ-reims.fr

structural dereplication of natural products.

Materials and Methods

All calculations were carried out on a DELL Precision 3530 laptop computer with 16 GB of RAM memory and an Intel® Core™ i5-8400H CPU @ 2.5 GHz running Windows 10 Education version 20H2. The ACD/C+H NMR Predictors and DB 2021.1.0 software was purchased from ACD/Labs (Toronto, Ontario, Canada). The library of cheminformatics functions RDKit version 2021.03.2 was run in the conda environment (https://www.anaconda.com) with python version 3.6.13 as programming language (https://www.python.org/). Additional chemical structure operations were carried out by functions implemented in the KnapsackSearch GitHub repository (https://github.com/nuzillard/KnapsackSearch).[8] The LOTUS database version 7 in file `220525_frozen_metadata.csv.gz`, was downloaded from the zenodo data repository, https://zenodo.org/record/6582124.

The unpacking of `220525_frozen_metadata.csv.gz` provided the text file `220525_frozen_metadata.csv` from which each line was scanned to constitute triplets made of the line number, of a wikidata identifier (QID, see https://www.wikidata.org) and of a SMILES chain.[9] There may be more than one line related to the same QID and SMILES; in this case only the first line was considered for future use. The text produced by the python script `csv2smi.py` from the triplets was redirected to file `lotusv7.smi`; it contained two columns, one for the SMILES chain and the other one for a compound character string formed by the QID and the line number joined by an underscore character, such as in Q43656_2, and used as compound name. Wikidata compound Q43656 is cholesterol and it appeared in file `z20525_frozen_metadata.csv` at line 2. The file `lotusv7.smi` contained a minimal description for 201022 compounds.

The script `smi2ACD.py` applied to `lotusv7.smi` resulted in file `fake_acd_lotusv7.sdf` in which a fake chemical shift value was assigned to each carbon atom (99.99). The resulting chemical shift lists, one per compound, were formatted to be understood by the ACD/Labs software as if they were experimental chemical shifts. File `fake_acd_lotusv7.sdf` contained 201022 compounds. Action of `tautomer.py` and `rdcharge.py` scripts achieved in–place tautomer correction and charge/valence correction. The necessity of using these scripts was reported in a preceding work.[3] Operations on structure files make use of the RDKit cheminformatics function library for SMILES chain interpretation, 2D structure diagram generation, tautomer correction and SDF structure file handling.

The `splitter.py` script applied to file `fake_acd_lotusv7.sdf` created 21 .sdf files stored in a dedicated subdirectory. These pieces were named `fake_acd_lotusv7_`xx`.sdf` with xx ranging from 00 to 20, the 20 first ones contained 10,000 compounds each and the last one contained the remaining structures. Chemical shift prediction by ACD/Labs software using the validation algorithm was carried out on these sub-files. Each sub-file was imported in an ACD/Labs database, validated for chemical shifts and exported as `fake_acd_lotusv7_`xx`_exported.sdf`. Action of script `CNMR_predict.py` on the latter produced the files `true_acd_lotusv7_`xx`.sdf` in which initial fake chemical shift values were replaced by predicted ones. The content of these files was concatenated into a single file, `acd_lotusv7_tmp.sdf`, containing 201,010 compound descriptions. The `supplement.py` script collected chemical taxonomy data present in file `220525_frozen_metadata.csv` and appended them to compound metadata in file `acd_lotusv7_tmp.sdf` in order to produce file `acd_lotusv7.sdf`. Chemical taxonomy data were obtained by the authors of LOTUS from NPclassifier and Classyfire.[10,11] The chemical identifiers SMILES, InChI and InChIKey[12] were recalculated from the structures in `acd_lotusv7.sdf` and the resulting InChIKeys compared to those provided by `220525_frozen_metadata.csv`. File `acd_lotusv7.sdf` was finally imported in the ACD database file `acd_lotusv7.NMRUDB` and in nmrshiftdb2 for structural dereplication works.

In addition, the prediction of the ^{13}C chemical shifts for the structures in `acd_lotusv7.sdf` was also carried out using the HOSE code approach. This was achieved by means of nmrshiftdb2 data from August 1st, 2019 and of the extended HOSE code search method described in[13]. The resulting predictions were also imported in nmrshiftdb2 to probe the influence of the predictor on dereplication.

Three natural product structures and their corresponding list of experimental chemical shift values were selected from published literature in order to illustrate the use of the acd_lotusv7 database for structural dereplication from ^{13}C NMR data. The used search engine was the one of nmrshiftdb2. The nmrshiftdb2 software supports the handling of experimental and predicted NMR spectra which can originate from various sources identified by category tags. The new category *acd_lotusv7* was introduced at the time of `acd_lotusv7.sdf` file importation. Another new category, *hose_lotusv7* was created to tag predictions produced by the HOSE code approach. The results of searches carried out with different selection criteria may be combined in nmrshiftdb2, including those by category tag, chemical shift list, and molecular formula. The nmrshiftdb2 spectrum search algorithm determines the similarity of two spectra by calculating a distance value. Therefore, the search results do not depend on a distance threshold or on any other parameter but are presented to the user after having been sorted in the order of decreasing similarity.

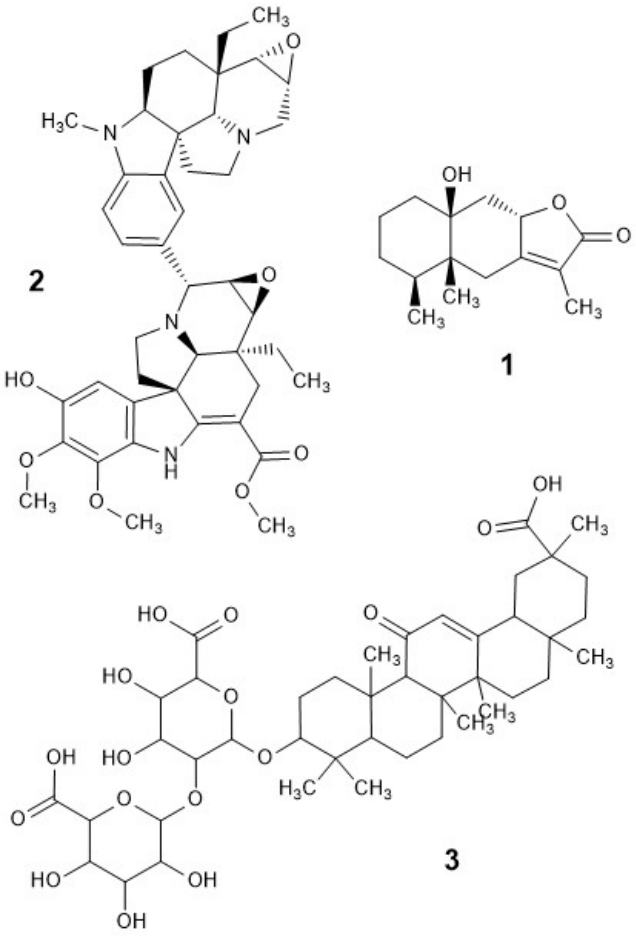
Results and Discussion

All the structures reported in LOTUSv7 file `220525_frozen_metadata.csv` but 12 were present in `acd_lotusv7.sdf`, thus constituting a very high success rate due to the careful design of the LOTUS database. However, compounds Q105187773 and Q105007277 were manually discarded as they erroneously contained divalent helium atoms. The other discarded compounds were left out by the ACD/Labs software at the time of `fake_acd_lotusv7_`xx`.sdf` file importation. Discrepancies between initial InChIKey descriptors and recalculated ones arose in 4167 compounds and originated from wrong writing or reading of configuration markers in 2D structure drawings.

The overall chemical shift prediction by the ACD/Labs sofware took about 30 hours, corresponding to an average processing rate of 2 compounds per second. The size of the final `acd_lotusv7.sdf` file was 1,256 MB. It can be read by any cheminformatics toolkit and its compressed version, of size 219 MB, was stored in a public zenodo repository at https://zenodo.org/record/6621129.

In the following, we illustrate the dereplication process by the retrieval of the structure of three natural compounds of

low, medium, and high complexity from published lists of carbon-13 NMR chemical shifts.



Scheme 1. Structure of compounds **1**, **2**, and **3**.

Compound **1** is characterized by a list of 15 ^{13}C NMR chemical shifts: 36.4, 22.3, 29.7, 33.5, 44.9, 31.8, 161.4, 78.7, 41, 75, 120.5, 176, 8.3, 14.7, and 16 ppm.[14] Searching in the *acd_lotusv7* category of nmrshiftdb2 was carried out with these values as chemical shift targets. The structure ranked first in the result set did fit will all constraints on chemical shift values and was identical to the published structure. The analyzed compound was an eremophilane sesquiterpene lactone isolated from *Hertia cheirifolia*, an Asteraceae collected in the North of Africa.[14] Searching in the whole of nmrshiftdb2 gave the same structure as the top hit, but showed some additional relatively good hits not found within the *acd_lotusv7* category. We also combined the search with a chemical formula search for $C_{15}H_{22}O_3$, the chemical formula of the top hit. This gave the same structure as the top hit, as expected, but excluded some hits from the top 10, which had a different elemental composition. This aspect of the corresponding search interface is shown in Figure 1. It should be noted that some of the top 10 are stereoisomers of the top hit. Since the top hit was the expected stereoisomer, the search has identified the correct compound. Wikidata links were added to the records in nmrshiftdb2 (visible on the Additional Data tab), so following the link for the top hit, Q105172965, lead to a list of three plants in which it was found. Selecting *Hertia cheirifolia* led to the reference to the original publication, referenced Q57818186 in Wikidata, and then to its DOI: 10.1016/0031-9422(90)83039-4, thus providing an incredibly quick way of establishing a relationship between a list of chemical shifts and an article published in a phytochemistry journal in 1990.

The same search was attempted using HOSE code prediction. The best match was the same as with the ACD/Labs prediction, but its similarity score was significantly lower (see Figure 2), mainly due to three shift values being almost 4 ppm away from the ACD/Labs prediction. The other top hits were also the same as with the ACD/Labs prediction as they all presented similar shift differences so that hit ranking was preserved.

Compound **2** was reported with 43 ^{13}C NMR chemical shifts: 165.1, 58.2, 47.7, 42.1, 54.5, 135.5, 104.0, 143.7, 138.6, 136.7, 128.2, 54.3, 56.5, 90.4, 23.5, 7.3, 26.8, 36.5, 61.6, 168.6, 60.8, 60.3, 50.8, 73.2, 52.7, 53.5, 40.8, 51.2, 136.5, 122.8, 122.4, 128.0, 105.9, 149.7, 52.6, 57.0, 19.7, 23.6, 7.4, 27.9, 34.5, 67.2, and 31.4 ppm.[15] Searching acd_lotusv7 under the same conditions as those used for compound **1**, produced stereoisomers and compounds with various formulae in the list of the ten best hits. Restricting the search to molecular formula $C_{43}H_{52}N_4O_7$ associated the correct structure to the list of chemical shift values in the original publication. The molecular formula of the second best hit was $C_{43}H_{54}N_2O_8$ and corresponded to the best hit compound but with a water molecule added, resulting from the opening of an epoxide ring. The best hit compound was isolated from *Ervatamia peduncularis* whose QID is Q96375411. Searching this QID in Wikidata revealed that this compound was named conofoline while it was published in 1995 as pedunculin by one of the authors.[16] Conofiline was isolated from *Tabernaemontana divaricata* and published shortly before pedunculin. It appeared that *Ervatamia peduncularis* and *Tabernaemontana peduncularis* are synonyms, suggesting that the two research groups studied two plants of the same genus, differing by their species name, but producing the same highly complex dimeric indolomonoterpenic alkaloids.

The similarity score of conofoline was only 35.09% with HOSE code prediction, whereas it reached 76.4% with the one by the ACD/Labs software. As for compound **1**, the lower score arose from a few predicted values being significantly off. The fifth hit from the ACD/Labs prediction, QID Q104944410, has a similarity score of 43.17% using the HOSE code prediction. Structure Q104944410 would therefore come first in a search using HOSE code predictions only.

Results in nmrshiftdb2 searches can potentially be improved by associating chemical shifts with carbon multiplicity information. The search tool of the ACD/Labs software does not offer this possibility. Multiplicity characterises methine, methylene, methyl, and quaternary carbon atoms and is reported in queries by one–letter symbols S(inglet), D(oublet), T(riplet), and Q(uadruplet), respectively. This information is experimentally provided either by 1D ^{13}C DEPT spectra or by multiplicity–edited 2D ^{1}H–^{13}C HSQC spectra. Searching for 165.1S, 58.2D, 47.7T, 42.1T, 54.5S, 135.5S, 104.0D, 143.7S, 138.6S, 136.7S, 128.2S, 54.3T, 56.5D, 90.4S, 23.5T, 7.3Q, 26.8T, 36.5S, 61.6D, 168.6S, 60.8Q, 60.3Q, 50.8Q, 73.2D, 52.7T, 53.5D, 40.8T, 51.2S, 136.5S, 122.8S, 122.4D, 128.0D, 105.9D, 149.7S, 52.6D, 57.0D, 19.7T, 23.6T, 7.4Q, 27.9T, 34.5S, 67.2D, and 31.4Q did not change the order of the previous structure

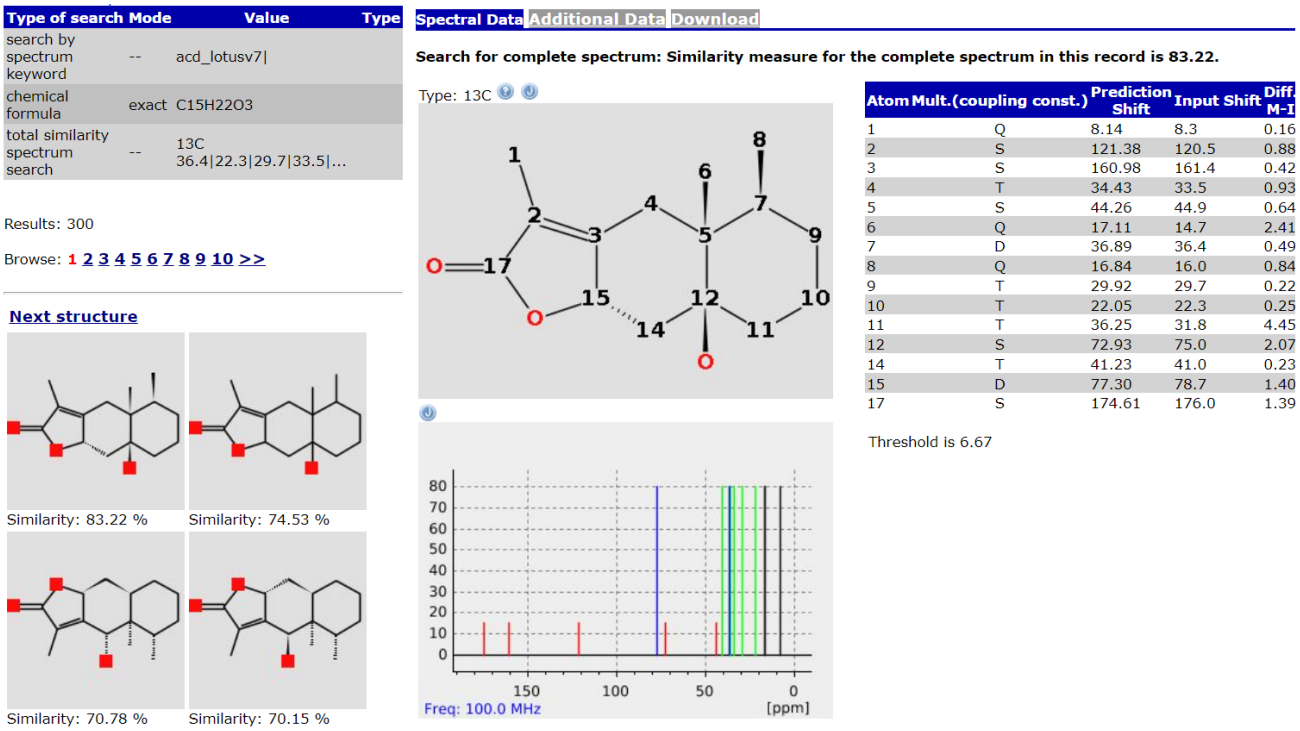


Figure 1. A search for compound **1**, restricted to category *acd_lotusv7* and formula $C_{15}H_{22}O_3$.

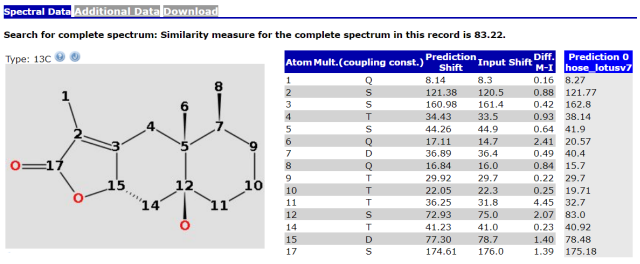


Figure 2. The ACD/Labs prediction (column Prediction) with the searched experimental shifts (column Input Shift) and the HOSE code prediction (column Prediction *hose_lotusv7*. The difference in predicted values for atoms 4, 7, and 11 was almost 4 ppm.

proposals, but the similarity scores for the HOSE code predicted spectra changed: 30.7% for conofoline, close to what is was without the multiplicty information (35.09%), whereas it decreased significantly from 43.17% to 33.93% for structure Q104944410. Even though introducing multiplicity did not produce the expected result when using the chemical shift prediction based on HOSE codes, it made the score of an incorrect best hit a lot less convincing than chemical shifts alone, thus enforcing the idea according to which more information helps to take better decisions.

Compound **3** was reported with 42 ^{13}C NMR chemical shifts: 39.6, 26.7, 89.3, 40.0, 55.6, 17.7, 33.1, 45.6, 62.2, 37.4, 199.4, 128.8, 169.4, 43.6, 26.9, 26.7, 32.2, 48.8, 41.8, 44.1, 31.7, 38.5, 28.2, 16.7, 16.9, 18.9, 23.6, 28.7, 28.8, 179.1, 105.1, 84.5, 77.8, 73.0, 77.3, 172.3, 106.9, 76.8, 77.7, 73.3, 78.4, and 172.0 ppm.[17] Searching the *acd_lotusv7* category in nmrshiftdb2 for this spectrum, a compound with 42 carbons was ranked first. The best hit, referenced Q105155240 in Wikidata, was associated with "liquorice" in this database and with the chemical study of *Glycyrrhiza uralensis*. The publication from which the ^{13}C NMR data were taken referred to *Glycyrrhiza glabra*, or sweet liquorice, and to *Glycyrrhiza uralensis*.[17] The corresponding compound is called glycyrrhizic acid, also known as glycyrrhizin. Its representation as a "flat" molecule is due to lack of configuration data in the original document that was imported in LOTUS. The best hit which had the same formula when searching for the spectrum without a formula restriction belonged to the class of triterpene saponins, as indicated by Classyfire data. This is the same class as for the expected structure.

A few other compounds were tested for dereplication from their ^{13}C NMR data but their number is certainly too small yet in order to suggest that acd_lotusv7 could be the definitive solution (if any exists) to the problem of the quick identification of already known natural products. However, the first trials were highly encouraging, mainly due to the richness of the content of the LOTUS database and to the quality of chemical shift prediction. No attempt were made to date to identify compounds from partial lists of chemical shifts by means of the acd_lotusv7 database. Reducing the completeness of input data results in less accurate and more diverse structure proposals, as already noticed during the course of many previous dereplication studies based on the CaraMel workflow.[1]

Conclusion

This article reports the building process of the acd_lotusv7 database using the natural product structures contained in the LOTUS database, version 7, and using ^{13}C NMR chemical shift predictions provided by the ACD/Labs predictor in the validation mode. Database users can carry

out structural dereplication either through the search tool of the ACD/Labs software or through the freely accessible interface of the nmrshiftdb2 web site. The latter offers as supplementary feature the possibility of associating the highly useful multiplicity information to chemical shift values in order to retain only the most NMR relevant candidate structures.[6] Database usage is illustrated by examples of various structural complexity. We also show that the quality of the prediction plays a significant role in the success of dereplication. Future works will involve the creation of a protocol intended to minimize the amount of calculation required by the updating for the acd_lotus database subsequently to the publication of new LOTUS versions.

Acknowledgements

JMN thanks the Centre National de la Recherche Scientifique for personal financial support.

Conflict of Interest

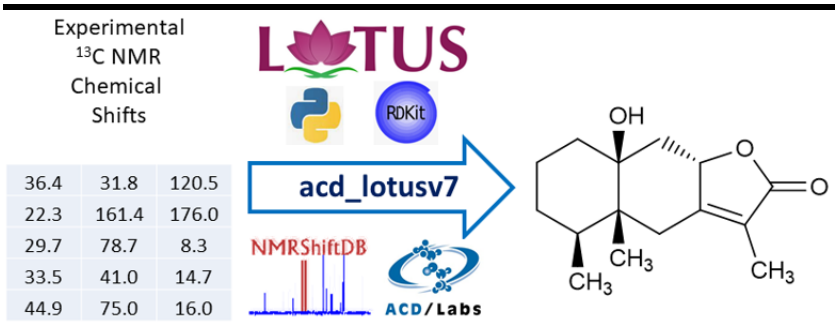
The authors have no conflicts of interest to declare.

Keywords: Natural Products • Nuclear Magnetic Resonance • Dereplication • Databases • Software

References

[1] J. Hubert, J.-M. Nuzillard, J.-H. Renault, *Phytochem Rev* **2017**, *16*, 55.
[2] J. Hubert, J.-M. Nuzillard, S. Purson, M. Hamzaoui, N. Borie, R. Reynaud, J.-H. Renault, *Anal Chem* **2014**, *86*, 2955, pMID: 24555703.
[3] J.-M. Nuzillard, *Analytica* **2021**, *2*, 50.
[4] A. Rutz, M. Sorokina, J. Galgonek, D. Mietchen, E. Willighagen, A. Gaudry, J. G. Graham, R. Stephan, R. Page, J. Vondrášek, C. Steinbeck, G. F. Pauli, J.-L. Wolfender, J. Bisson, P.-M. Allard, *eLife* **2022**, *11*, e70780.
[5] S. Kuhn, N. E. Schlörer, *Magn Reson Chem* **2015**, *53*, 4263.
[6] A. Bruguière, S. Derbré, C. Coste, M. Le Bot, B. Siegler, S. T. Leong, S. N. Sulaiman, K. Awang, P. Richomme, *Fitoterapia* **2018**, *131*, 59.
[7] M. Sorokina, C. Steinbeck, *J Cheminform* **2020**, *12*, 20.
[8] M. Lianza, R. Leroy, C. Machado Rodrigues, N. Borie, C. Sayagh, S. Remy, S. Kuhn, J.-H. Renault, J.-M. Nuzillard, *Molecules* **2021**, *26*.
[9] D. Weininger, *J Chem Inf Comput Sci* **1988**, *28*, 31.
[10] H. W. Kim, M. Wang, C. A. Leber, L.-F. Nothias, R. Reher, K. B. Kang, J. J. J. van der Hooft, P. C. Dorrestein, W. H. Gerwick, G. W. Cottrell, *J Nat Prod* **2021**, *84*, 2795, pMID: 34662515.
[11] Y. Djoumbou Feunang, R. Eisner, C. Knox, L. Chepelev, J. Hastings, G. Owen, E. Fahy, C. Steinbeck, S. Subramanian, E. Bolton, R. Greiner, D. S. Wishart, *J Cheminform* **2016**, *8*, 61.
[12] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, D. Tchekhovskoi, *J Cheminform* **2015**, *7*, 23.
[13] S. Kuhn, S. R. Johnson, *ACS Omega* **2019**, *4*, 7323, pMID: 31459832.
[14] G. Massiot, J.-M. Nuzillard, L. Le Men-Olivier, P. Aclinou, A. Benkouider, A. Khelifa, *Phytochemistry* **1990**, *29*, 2207.
[15] M. Zèches-Hanrot, J.-M. Nuzillard, B. Richard, H. Schaller, H. A. Hadi, T. Sévenet, L. L. Men-Olivier, *Phytochemistry* **1995**, *40*, 587.
[16] T.-S. Kam, S. Anuradha, *Phytochemistry* **1995**, *40*, 313.
[17] L. A. Baltina, O. Kunert, A. A. Fatykhov, R. M. Kondratenko, L. V. Spirikhin, L. A. Baltina Jr., F. Z. Galin, G. A. Tolstikov, E. Haslinger, *Chem Nat Compd* **2005**, *41*, 432.

Entry for the Table of Contents



The quick identification of known low molecular weight organic compounds of natural origin by means of ^{13}C NMR data is made easy by the querying of the acd_lotusv7 database that includes structures from the LOTUS database and ACD/Labs chemical shift predictions. The open source nmrshiftdb2 web interface at https://nmrshiftdb.nmr.uni-koeln.de provides a simple and free way to retrieve compounds from acd_lotusv7 and to analyze search results.