Structural dereplication of natural products by means of carbon-13 nuclear magnetic resonance data.

Jean-Marc Nuzillard

▶ **To cite this version:**

Jean-Marc Nuzillard. Structural dereplication of natural products by means of carbon-13 nuclear magnetic resonance data.. 2022. hal-03704922

HAL Id: hal-03704922
https://hal.univ-reims.fr/hal-03704922

Preprint submitted on 26 Jun 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Structural dereplication of natural products by means of carbon-13 nuclear magnetic resonance data.

Author: J-M Nuzillard (ORCID : 0000-0002-5120-2556)

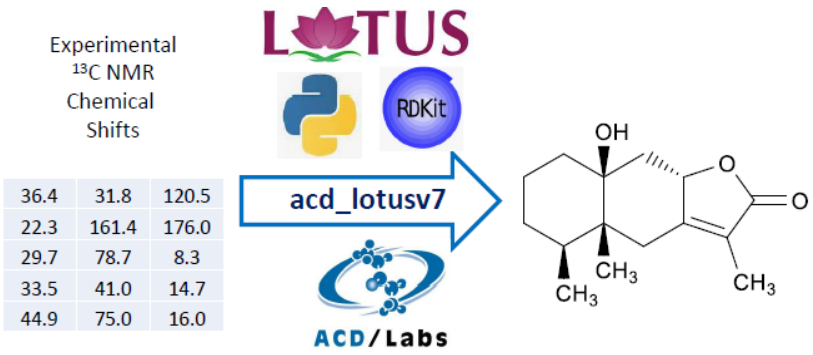
Correspondence: jm.nuzillard@univ-reims.fr

Affiliation: Université de Reims Champagne Ardenne, CNRS, ICMR UMR 7312, 51097 Reims, France

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

Abstract: The present article reports the creation and usage of a general natural product database for structural dereplication of natural products. This database, acd_lotusv7, is based on the LOTUS natural products database as the unique source of chemical structures. Database construction and use for dereplication relies on the commercial ACD/Labs C+H Predictor and DB software. The linkage of each natural compound with a Wikidata resource identifier accelerates the access to the primary literature data such as biologic origin and bibliographic references.

Graphical abstract:



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 organic natural product chemistry, a compound produced by different living organisms may have been purified many times by different research teams resulting in repeated structure elucidation works. 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 combines purification and structural dereplication in a single workflow called CaraMel.[1] With this procedure, the isolation to a high purity level of all the constituent of a mixture 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 ^{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 successively 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) called C+H Predictors and DB (hereafter referred to as the ACD/Labs software, even though the ACD/Labs company provides many other software products).[2]

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 user supplied NMR chemical shifts values assigned to used supplied molecular structures. This predictor (hereafter, 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 to update the initially given structure assignments so that the predicted ones appear as if they were experimental. This way of supplementing a collection of structures by means of the ACD/Labs validation algorithm was already proposed in the framework of the constitution 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, is based on the LOTUS natural products database.[4] The choice of ^{13}C NMR spectroscopy as primary identification data source and the resorting on predicted chemical shift values have already been discussed in previous works.[1,5]

Material 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 (anaconda.com) with python 3.6.13 as programming language.[6] Additional chemical structure operations were carried out by functions implemented in the KnapsackSearch GitHub repository (https://github.com/nuzillard/KnapsackSearch).[7] The LOTUS database version 7 in file 220525_frozen_metadata.csv.gz, was downloaded from the zenodo data repository, https://zenodo.org/record/6582124.[8]

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 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 appears in file 220525_frozen_metadata.csv at line 2. The file lotusv7.smi contains 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 contains 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 sub-directory. 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 the script CNMR_predict 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 assembled 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 for future structural dereplication works.

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.

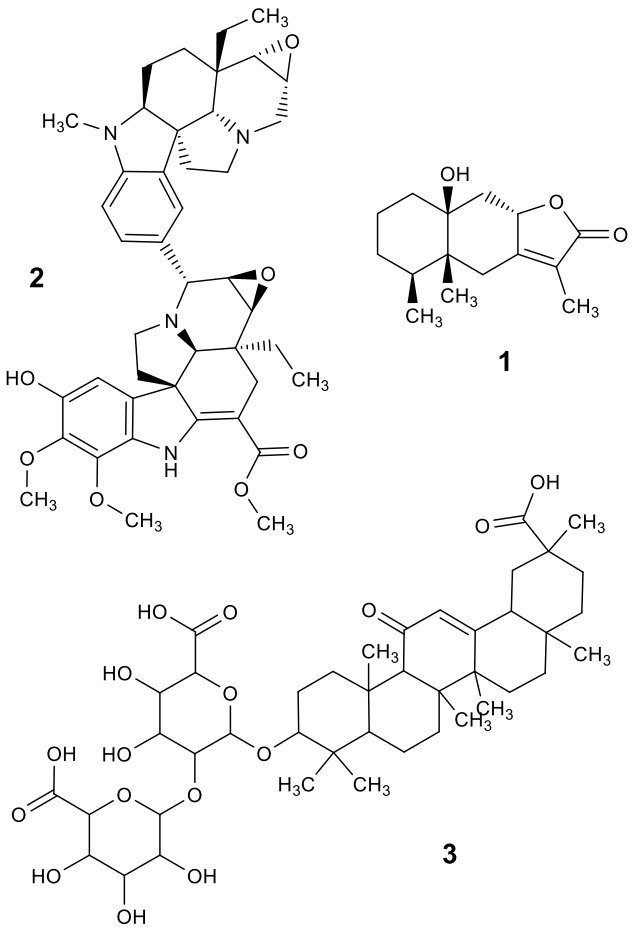
Results

All the structures reported in LOTUSv7 file 220525_frozen_metadata.csv but 12 were imported in acd_lotusv7.NMRUDB, 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 were erroneously supposed to contain 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 configurational features in 2D structure drawings.

The overall chemical shift prediction task took about 30 hours, corresponding to an average processing rate of 2 compounds per second. The size of the final acd_lotusv7.NMRUDB file was 706 MB. The acd_lotusv7.sdf file can be read by any cheminformatics tool and its compressed version, of size 219 MB, was stored in a zenodo repository at https://zenodo.org/record/6621129 .

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.[13] Searching in database acd_lotusv7 was carried out with these values as chemical shift targets, requesting at least the matching of 15 chemical shift values with a looseness of 3 ppm and solution ranking based on minimal distances. The list of results contained 177 structures, among which the one ranked first did fit will all constraints on chemical shift values and was identical to the published structure, the one of an eremophilane sesquiterpene isolated in our group from *Hertia cheirifolia*, an Asteraceae collected in the North of Africa.[13] Searching

Wikidata for the identifier related to the selected compound, Q105172965, lead to list of three plants in which it was found. Selecting *Hertia cheirifolia* led to the reference of the original publication, referenced Q57818186 in Wikidata, and then to the 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. This structure is the first one of a natural product that was determined in 1989 by the author of the present article, when there were no computers in most of academic chemistry laboratories and at a time this story would have been considered as pure science-fiction.



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

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.[14] Searching acd_lotusv7 in the same conditions as those used for compound **1**, requesting at least 43 chemical shift matches lead to 3 hits. Only the first one accounted for the molecular formula $C_{43}H_{52}N_4O_7$ associated to the list of chemical shift values in the original publication. The molecular formula of the two other hits was $C_{43}H_{54}N_2O_8$ and corresponded to the first hit compound to which of a water molecule was added, resulting from the opening of an epoxide ring. The first hit corresponded to a compound isolated from *Ervatamia peduncularis* whose QID is Q96375411. Searching this QID in Wikidata revealed that this compound was named conofoline while our group published it in 1995 as pedunculin.[15] 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 alkaloid.

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.[16] Searching acd_lotus_v7 for at least 42 chemical shift matches resulted in 2 hits, a compound with 42 carbons, ranked first, and another one with 41 carbons but for which only 38 chemical shift matched. These compounds belong to the family of triterpene saponins, as indicated by the to classyfire data embedded in acd_lotusv7. The best hit, referenced Q105155240 in Wikidata, was associated to "liquorice" in this database and to 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*.[16] 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.

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 ensure that the acd_lotusv7 would 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 on the quality of chemical shift prediction.

Supplementary material

The acd_lotusv7.sdf file, ready for importation in the ACD/Labs software in available from https://zenodo.org/record/6621129 .

The software tools for the construction of the acd_lotusv7.sdf file are available from https://github.com/nuzillard/KnapsackSearch/tree/master/acd_lotusv7_add_on .

Conclusion

This article reports the building process of the acd_lotusv7 database using the structural data 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 usage is illustrated by examples of various structural complexity. Future works may 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. The actual structure search engine, based on the ACD/Labs software, does not take into account the highly useful multiplicity information bound to ^{13}C NMR resonances.[5] A remedy to this situation will be necessary to propose to potential users, as well as an integrated graphical interface for interactive solution structure browsing and evaluation.

References.

1. Identification of Natural Metabolites in Mixture: A Pattern Recognition Strategy Based on ^{13}C NMR. J Hubert, J-M Nuzillard, S Purson, M Hamzaoui, N Borie, R Reynaud, J-H Renault; Anal. Chem. 86 (2014) 2955−2962; DOI : 10.1021/ac403223f

2. https://www.acdlabs.com/

3. Taxonomy-Focused Natural Product Databases for Carbon-13 NMR-Based Dereplication. J-M Nuzillard; Analytica 2 (2021) 50–56; DOI: 10.3390/analytica2030006

4. The LOTUS initiative for open knowledge management in natural products research. 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 11 (2022) e70780; DOI: 10.7554/eLife.70780

5. ^{13}C NMR Dereplication of Garcinia extracts: Predicted Chemical Shifts as reliable Databases. A Bruguière, ·S Derbré, C Coste, M Le Bot, B Siegler, S T Leong, S N Sulaiman, K Awang, P Richomme; Fitoterapia 2018, 131, 59–64; DOI: 10.1016/j.fitote.2018.10.003

6. www.rdkit.org

7.The Three Pillars of Natural Product Dereplication. Alkaloids from the Bulbs of Urceolina peruviana (C. Presl) J.F. Macbr. as a Preliminary Test Case. M Lianza, R Leroy, C Machado Rodrigues, N Borie, C Sayagh, S Remy, S Kuhn, J-H Renault, J-M Nuzillard; Molecules 26 (2021) 637; DOI: 10.3390/molecules26030637

8. www.zenodo.org

9. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. D Weininger; J. Chem. Inf. Comput. Sci. 28 (1988) 31–36; DOI: 10.1021/ci00057a005

10. NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. 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. 84 (2021) 2795–2807; DOI: 10.1021/acs.jnatprod.1c00399

11. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. Y D 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. Cheminformatics 8 (2016) 61; DOI: 10.1186/s13321-016-0174-y

12. InChI, the IUPAC International Chemical Identifier. S R Heller, A McNaught, I Pletnev, S Stein, D Tchekhovskoi; J. Cheminformatics 7 (2015) 23; DOI: 10.1186/s13321-015-0068-4

13. Eremophilenolides from Hertia cheirifolia. G Massiot, J-M Nuzillard, L Le Men-Olivier, P Aclinou, A Benkouider, A Khelifa; Phytochemistry 29 (1990) 2207-2210; DOI: 10.1016/0031-9422(90)83039-4

14. Alkaloids from leaves and stem bark of Ervatamia peduncularis. M Zèches-Hanrot, J-M Nuzillard, B Richard, H Schaller, H A Hadi, T Sévenet, L Le Men-Olivier; Phytochemistry 40 (1995) 587-591; DOI: 10.1016/0031-9422(95)00152-W

15. Alkaloids from Tabernaemontana divaricata. T-S Kam, S Anuradha; Phytochemistry 40 (1995) 313-316; DOI: 10.1016/0031-9422(95)00266-A

16. High-Resolution ^{1}H and ^{13}C NMR of Glycyrrhizic Acid and Its Esters. 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. 41 (2005) 432–435; DOI: 10.1007/s10600-005-0171-2