doi:10.20944/preprints202105.0701.v1

Taxonomy-focused natural product databases for carbon-13 NMR-based dereplication

Jean-Marc Nuzillard

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

jm.nuzillard@univ-reims.fr

Abstract: The recent revival of the study of organic natural products as renewable sources of medicinal drugs, cosmetics, dyes, and materials motivated the creation of general-purpose structural databases. Dereplication, the efficient identification of already reported compounds, relies on the grouping of structural, taxonomic and spectroscopic databases that focus on a particular taxon (species, genus, family, order…). A set of freely available python scripts, CNMRPredict, is proposed for the quick supplementation of taxon-oriented search results from the LOTUS database (lotus.naturalproducts.net) with predicted carbon-13 NMR data from the ACD/Labs (acdlabs.com) CNMR predictor and DB software to provide easily searchable databases. The database construction process is illustrated using *Brassica rapa* as taxon example.

Keywords: Natural products, databases, dereplication, taxonomy, NMR

Main text

Dereplication in the context of natural product (NP) chemistry may be defined as the identification of known chemotypes, so that structure re-elucidation and possibly compound re-isolation can be avoided.[1, 2] Establishing whether an organic compound is known requires the availability of a collection of identity cards of known compounds, possibly organized as a computer database (DB). The existence, availability, scope, and limitations of the numerous NP DBs has been thoroughly reviewed recently,[3] resulting in the creation of a new DB named COCONUT (coconut.naturalproducts.net) in which the content of numerous DBs was collected.[4] An even more recent work led to LOTUS (lotus.naturalproducts.net), a DB that connects NP molecular structures with the taxonomic classification of the organisms they originate from and that constitutes an utterly useful source of data for NP chemistry.[5] Moreover, the LOTUS database provides bibliographic links to compound descriptions. The motivation for the renewed interest toward NP studies arises from their ability to propose highly diverse and renewable sources of medicinal drugs, cosmetics, dyes, and materials in the broader sense.

NP chemical studies start from taxonomically well-defined biological resources from which the products of the metabolism, primary and specialized, is extracted. Extraction is a science in itself, it has considerably evolved during the last decades, involving a wide range of physical and chemical processes adapted to the nature of the starting material and to the desired extraction selectivity.[6] Crude NP extracts are generally substances made of highly complex compound mixtures. The reward of the subsequent extract complexity reduction by fractionation and purification is a simplification of the identification task. Alternatively, studying complex mixtures results in challenging identification problems but reduces the investment in separation techniques.

The hyphenation of liquid chromatography (LC) and mass spectrometry (MS) for extract analysis takes advantage of extremely powerful purification devices(UPLC chromatographers) with extremely sensitive detection devices (mass spectrometry, possibly with MS^n capabilities) so that the extract fractionation steps may be as reduced as possible. Compounds are identified from their exact molecular formula, fragmentation pattern, and ionic mobility. Fragmentation pattern analysis has proved to be highly successful and lead to initiatives such as GNPS, which results from collaborative efforts among numerous scientists.[7] LC-MS based methods frequently provide annotations rather than identification meaning that the collected experimental data may fit with isomers collections. Ideally, identification succeeds when an annotation set can be reduced to a single compound.

The use of LC hyphenated with nuclear magnetic resonance (NMR) spectroscopy is frequently limited by the amount of purified compound that can be analysed, NMR being far less sensitive than MS. Mixture analysis by NMR is a topic in itself and methods are available for the analysis of crude extracts or for series of extract fractions.[8-11] NMR characterizes molecular compounds at the atomic level so that NMR experimental data are less prone than MS data to be compatible with a high number of molecular structures. Ambiguity from NMR arise often from the lack of configuration assignment at chiral structure elements while planar structures are generally defined to a high level of accuracy.[12]

Dereplication relies on the comparison between freshly collected spectroscopic data with those from previous studies and stored in a DB. The extraction of experimental MS and NMR from published data is a tedious process that may result in copy errors and in the exact copy of erroneously structure or data assignments. However, the accumulated knowledge gained on the relationships between molecular structures and measurement outcomes has made possible to design spectroscopic prediction tools that may replace, to some extent, experimental spectral data by predicted ones. [8, 13]

The analytical technique being either MS or NMR or both, dereplication of NPs relies on NP DBs containing structural, taxonomic and spectroscopic data.[14] Merging spectroscopic and biological taxonomy data offers a way to reduce, possibly to one, the number of annotations for a given compound. Restricting the set of candidate structures for dereplication to the chemical entities produced by the organisms that are taxonomically related to the one under study finds its justification in the co-evolution of species and of the compounds they produce to establish relationships with their environment. This Communication reports a way to create a database related to a given taxon with included data for dereplication through ^{13}C NMR spectroscopy. A similar approach, KnapsackSearch, was reported, resorting on the internet access to the KNApSAcK DB (knapsackfamily.com).[14] The current approach, called CNMRPredict, relies on the LOTUS DB as structure provider by the possibility it offers to carry out searches according to taxonomy and to easily export the result of searches.[5] CNMRPredict also relies on the Advanced Chemistry Development, Inc. (ACD/Labs) CNMR Predictor and DB software (version 2020.1.0) for high-quality ^{13}C NMR chemical shift prediction and for easy structure search (acdlabs.com) and on the python version of the RDKit cheminformatics library (rdkit.org, version 2021.03.2). The source code of the python scripts related to CNMRPredict are freely available from GitHub (github.com/nuzillard/KnapsackSearch/).

The creation of a focused library is illustrated here for turnip, or *Brassica rapa*. Using this binomial species name as a simple search key in the LOTUS DB through its web interface results in 121 hits. The search result is downloaded as file lotus_simple_search_result.sdf in V3000 SDF format (www.daylight.com/meetings/mug05/Kappler/ctfile.pdf) and stored in a local computer directory, in which all files related to the turnip project are stored. A new ACD/Labs DB, turnip.NMRUDB, is created and filled with data from lotus_simple_search_result.sdf. Exporting this database in SDF format as file turnip.sdf has its conversion to the V2000 SDF format as a side-effect.

The turnip.sdf file may contain identical structures, apparently because different InChI character strings in LOTUS may result in the production of structures in lotus_simple_search_result.sdf that are in turn recoded as identical InChI strings by the RDKit library.[15] A python script, uniqInChI.py, keeps only a single occurrence of duplicated structures according to InChI equality and is applied to file turnip.sdf, in which only one compound over 121 is removed.

Many structures downloaded from LOTUS were produced by the decoding of InChI strings. This process has the very visible side effect of replacing secondary and primary amide functions by their tautomeric iminol forms. Because the central carbon atom in enamine and iminol functional groups have their ^{13}C NMR chemical shift values not identically predicted, it appeared to be necessary to transform aliphatic iminols into their amide tautomer, as achieved by the tautomer.py script applied to file turnip.sdf. It should be noticed that the systematic (iupac.org) nomenclature of iminol-containing compounds in LOTUS is determined as if they were really iminols and not amides, resulting for example in a difficult identification of peptidic bonds in peptides.

Script tautomer.py relies on RDKit to write SDF files and makes internally use of reaction SMARTS (daylight.com/dayhtml/doc/theory/theory.smarts.html). Electrically charged atoms in structures written by RDKit include a non-default specification for the non-standard valence of such atoms (such as 4 for the nitrogen atom in an ammonium group), in accordance with SDF specification. Such a structure description is not interpreted as expected by the ACD/Labs software, thus precluding the prediction of chemical shifts. The rdcharge.py script resets the valence data piece to the default, non-blocking value and is applied to turnip.sdf.

The first step toward the automatic calculation of ^{13}C NMR chemical shift is to let the ACD/Labs software consider that experimental values were stored in an SDF file it produced, something feasible by supplementing the SDF file with data lines under the purposely created CNMR_SHIFTS SDF tag. These fake data lines include the fake chemical shift value 99.99, one per carbon atom in each molecule. The fakefakeACD.py script applied to turnip.sdf transforms it into file fake_acd_turnip.sdf.

For chemical shift prediction, a new ACD/Labs DB, fake_acd_turnip.NMRUDB, is created and filled by importation from file fake_acd_turnip.sdf. All carbon atoms appear with their arbitrarily given 99.99 chemical shift value. The presence of these values allows ACD/Labs DB to check all chemical shift values of all molecules from a single mouse click. Checking the chemical shifts of a DB that does not contain chemical values fails to give a meaningful result, thus justifying the resorting to the fakefakeACD.py script. Exporting the current DB as file fake_acd_turnip_exported.sdf first displays a message that warns that the calculated chemical shifts will not be exported. This is simultaneously true and false. This is true because the calculated values cannot be used for structure search according to chemical shift similarity between stored values and a set of targeted values, as required for dereplication. This is also false because the result of the prediction is stored in the resulting file, here fake_acd_turnip_exported.sdf, under the CNMR_CALC_SHIFTS SDF tag.

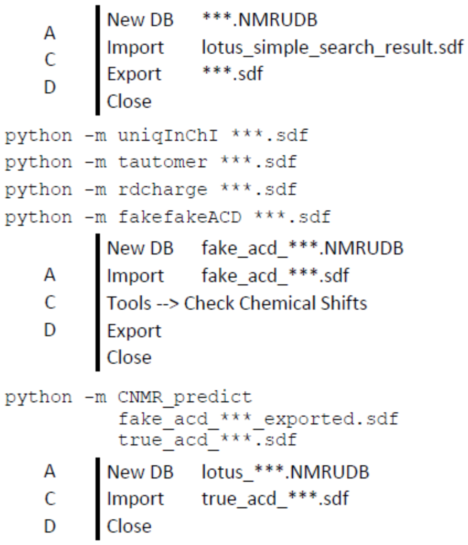
The next step toward a DB usable for dereplication consists in replacing the 99.99 values under the CNMR_SHIFTS SDF tag that are still present in file fake_acd_turnip_exported.sdf by the calculated values written under the CNMR_CALC_SHIFTS SDF tag. This operation is carried out by the script CNMR_predict.py acting on file fake_acd_turnip_exported.sdf to produce file

doi:10.20944/preprints202105.0701.v1

true_acd_turnip.sdf. A new ACD/Labs DB, lotus_turnip.NMRUDB is finally created and filled with compounds from file true_acd_turnip.sdf. The DB lotus_turnip.NMRUDB is now ready for compound search according to predicted ^{13}C NMR values. The file true_acd_turnip.sdf contains also SDF tags that make dereplication possible by the MixONat software. The script ACD_to_DerepCrude.py formats the predicted chemical shifts for its use with the DerepCrude software. Both DerepCrude[10] and MixONat[11] are dedicated to the dereplication by ^{13}C NMR either on crude NP extracts or on extract fractions, as alternatives to the now well-established CARAMEL dereplication procedure (nat-explore.com).[8]

The creation of DB lotus_turnip.NMRUDB is a process that alternates the execution of python scripts from a terminal window and the handling (create/import/predict/export/close) of ACD/Labs DB files. A template text file is proposed with the CNMRPredict project files so that the actions to perform sequentially can be easily accomplished. Figure 1 illustrates the content of this template file. CNMRPredict is a follow-up of the KnapsackSearch project that made use of nmrshiftdb2 for the prediction of ^{13}C NMR chemical shift values.[16] These predicted values may be formatted as experimental values under the CNMR_SHIFTS SDF tag and a template file is also available for this option.

Creating a file such as lotus_turnip.NMRUDB from the initial turnip.sdf without CNMRPredict would require a tedious compound-by-compound operation, lasting about one minute per structure. The prediction involving CNMRPredict lasts less than one second per structure, making it easy to use for the creation of taxonomically focused collections of natural products. For example, *Brassica rapa* belongs to the family of Brassicaceae and searching for this taxon in LOTUS results in 2271 compounds. A ready-to-search database of compounds from Brassicaceae can be thus produced in less than one hour on a standard laptop computer, an hour during which the computer is the only one that achieves the repetitive work.



```
       | New DB    ***.NMRUDB
   A   | Import    lotus_simple_search_result.sdf
   C   | Export    ***.sdf
   D   | Close
python -m uniqInChI ***.sdf
python -m tautomer ***.sdf
python -m rdcharge ***.sdf
python -m fakefakeACD ***.sdf
       | New DB    fake_acd_***.NMRUDB
   A   | Import    fake_acd_***.sdf
   C   | Tools --> Check Chemical Shifts
   D   | Export
       | Close
python -m CNMR_predict
          fake_acd_***_exported.sdf
          true_acd_***.sdf
   A   | New DB    lotus_***.NMRUDB
   C   | Import    true_acd_***.sdf
   D   | Close
```

Figure 1. Imaged view of the content of the template file that leads from a LOTUS search result file to an ACD/Labs database file with predicted ^{13}C NMR chemical shift values. The *** are intended to be replaced by some name, like “turnip” in the present example.

It should be noticed that DBs refer to published data and can propagate errors. For example, glucosinolates constitute an emblematic group of compounds related to the family of Brassicaceae (and more generally to the order of Capparales) that contain an *O*-sulfated anomeric (*Z*)-thiohydroximate function in which the double bond configuration may appear in DBs with the double bond in the (*E*) configuration or left undefined.[17] The library of the compounds from *Brassica rapa*, a Brassicaceae, reported by LOTUS contain such erroneous or incompletely defined structures that would be also found in reference databases such as the one of the Chemical Abstract Service (cas.org).

Briefly stated, the CNMRPredict project presented in this article allows one to easily and quickly combine structural and taxonomic data from the LOTUS NP database with ^{13}C NMR data predicted by the ACD/Labs CNMR Predictor in order to facilitate the ^{13}C NMR based dereplication of natural products. Future works will include the prediction of ^{1}H NMR chemical shifts and possibly of 2D NMR spectra.

References

[1] J. A. Beutler, A. B. Alvarado, D. E. Schaufelberger, P. Andrews, T. G. McCloud, *J. Nat. Prod.* **1990**, *53*, 867–874.

[2] J. Hubert, J.-M. Nuzillard, J.-H. Renault, *Phytochem. Rev.* **2017**, *16***,** 55–95.

[3] M. Sorokina, C. Steinbeck, *J. Cheminform.* **2020**, *12***,** 20.

[4] M. Sorokina, P. Merseburger, K. Rajan, *et al.*, *J. Cheminform.* **2021**, *13***,** 2.

[5] A. Rutz, M. Sorokina, J. Galgonek, D. Mietchen, E. Willighagen, J. Graham, R. Stephan, R. Page, J. Vondrášek, C. Steinbeck, G. F. Pauli, J.-L. Wolfender, J. Bisson, P.-M. Allard, *bioRxiv* **2021**.02.28.433265.

[6] C. Picot-Allain, M. F. Mahomoodally, G. Ak, G. Zengin, *Curr. Opin. Food Sci.* **2021**, *40*, 144–156.

[7] A. T. Aron, E. C. Gentry, K. L. McPhail, *et al.*, Nat. Protoc. **2020**, *15*, 1954–1991.

[8] J. Hubert, J.-M. Nuzillard, S. Purson, M. Hamzaoui, N. Borie, R. Reynaud, J.-H. Renault, *Anal. Chem.* **2014**, *86*, 2955–2962.

[9] J.-M. Nuzillard, P. Lameiras, *Prog. Nucl. Magn. Reson. Spectrosc.* **2021**, *123*, 1–50.

[10] A. Bakiri, J. Hubert, R. Reynaud, S. Lanthony, D. Harakat, J.-H. Renault, J.-M. Nuzillard, *J. Nat. Prod.* **2017**, *80*, 1387–1396.

[11] A. Bruguière, S. Derbré, J. Dietsch, J. Leguy, V. Rahier, Q. Pottier, D. Bréard, S. Suor-Cherer, G. Viault, A.-M. Le Ray, F. Saubion, P. Richomme, *Anal. Chem.* **2020**, *92*, 8793–8801.

[12] G. Lauro, G. Bifulco, *Eur. J. Org. Chem.* **2020**, 3929–3941.

[13] 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.

[14] M. Lianza, R. Leroy, C. Machado Rodrigues, N. Borie, C. Sayagh, S. Remy, S. Kuhn, J.-H. Renault, J.-M. Nuzillard, *Molecules* **2021**, *26*, 637.

[15] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, D. Tchekhovskoi, *J. Cheminform.* **2015**, *7*, 23.

[16] C. Steinbeck, S. Kuhn, *Phytochemistry* **2004**, *65*, 2711–2717.

[17] N. Ibrahim, I. Allart-Simon, G. R. De Nicola, R. Iori, J.-H. Renault, P. Rollin, J.-M. Nuzillard, *J. Nat. Prod.* **2018**, *81*, 323–334.