analytica

MDPI

Communication

Taxonomy-Focused Natural Product Databases for Carbon-13 NMR-Based Dereplication

Jean-Marc Nuzillard

ICMR UMR 7312, Université de Reims Champagne Ardenne, CNRS, 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, etc.). A set of freely available python scripts, CNMR_Predict, is proposed for the quick supplementation of taxon oriented search results from the naturaL prOducTs occUrrences database (LOTUS, lotus.naturalproducts.net) with predicted carbon-13 nuclear magnetic resonance data from the ACD/Labs CNMR predictor and DB software (acdlabs.com) to provide easily searchable databases. The database construction process is illustrated using *Brassica rapa* as a taxon example.

Keywords: natural products; databases; dereplication; taxonomy; NMR

Citation: Nuzillard, J.-M. Taxonomy-Focused Natural Product Databases for Carbon-13 NMR-Based Dereplication. *Analytica* **2021**, *2*, 50–56. https://doi.org/10.3390/analytica2030006

Academic Editor: Marc-Andre Delsuc

Received: 7 June 2021
Accepted: 25 June 2021
Published: 28 June 2021

1. Introduction

The motivation behind the renewed interest in natural product (NP) studies arises from their ability to propose highly diverse and renewable sources of medicinal drugs, cosmetics, dyes, and materials in the broader sense. Dereplication in the context of 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, resulting in the creation of a new DB named COCONUT in which the content of numerous DBs was collected [3–5]. An even more recent work led to LOTUS, a comprehensive fully open source open data DB that connects NP molecular structures with the taxonomic classification of the organisms they originate from in an unprecedented way and constitutes a significantly useful source of data for NP chemists [6,7]. Moreover, the LOTUS database provides bibliographic links to compound descriptions.

NP chemical studies start from taxonomically well-defined biological resources from which the products of the metabolism, primary and specialized, are extracted. Extraction 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 [8]. 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 led to initiatives such as Global Natural Products Social molecular networking (GNPS), which results from collaborative efforts among numerous scientists [9]. LC–MS based methods frequently provide annotations rather than identification, meaning that the collected experimental data may fit with isomer 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 analyzed, NMR being far less sensitive than MS. Methods have recently been made available for mixture analysis by NMR with applications to crude natural extracts or to series of extract fractions [10–13]. 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 arises often from the lack of configuration assignment to chiral structure elements, while planar structures are generally defined to a high level of accuracy [14]. While NMR spectra offer the possibility of distinguishing between diastereomers, even for compounds in mixtures, enantiomer identification requires either the isolation of pure compounds for their study by chiroptical methods or chemical derivatization by a homochiral reagent and subsequent NMR analysis [15,16].

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 data from publications is a tedious process that may result in copy errors and in the exact copy of erroneous structure or data assignments [17]. However, the accumulated knowledge gained of the relationships between molecular structures and measurement outcomes has made it possible to design spectroscopic prediction tools that may replace, to some extent, experimental spectral data by predicted ones [10,18].

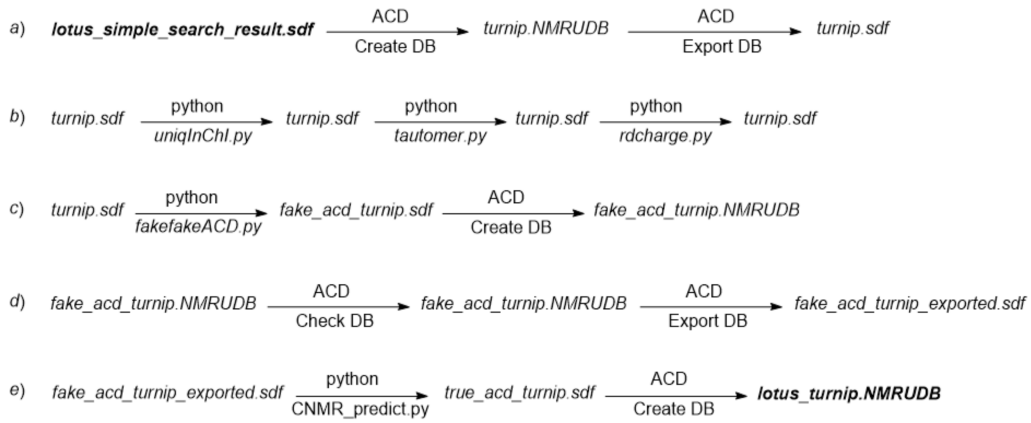
The dereplication of NPs relies on NP DBs containing structural, taxonomic and spectroscopic data [19]. Merging spectroscopic and biological taxonomy data offers a way to reduce, possibly to one, the number of annotations for a given compound. Restricting the set of candidate structures for dereplication to the chemical entities produced by the organisms that are taxonomically related to the one under study finds its justification in the coevolution 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 [19], was reported, resorting to the internet to access the KNApSAcK DB [20]. The current approach, called CNMR_Predict, relies on the LOTUS DB as a structure provider by the possibility it offers to carry out searches according to taxonomy and to easily export the result of searches [6]. The use of ^{13}C NMR data for dereplication seems inappropriate in this context, considering the higher sensitivity of ^{1}H NMR. The advantages of ^{13}C NMR lie in the ubiquitous presence of carbon atoms in organic molecules, in the absence of a signal fine structure, so that one carbon atom creates only one spectral peak, in the production of resonances that are narrow (about 1 Hz) by comparison to spectral widths (tens of kHz), resulting in a low probability of peak superimposition, in the low sensitivity of ^{13}C chemical shift to solvent and temperature effects, and in the accurate predictability of these chemical shifts [10]. These features make of ^{13}C NMR a useful tool for NP dereplication. A general review article about NP mixture analysis has been recently published [21]; it includes references to the methods that benefit from taxonomy focused NMR databases such as CARAMEL [10], DerepCrude [12], or MixONat [13], but in which the step of ^{13}C NMR chemical shift prediction still constitutes a bottleneck.

2. Materials and Methods

The LOTUS DB was searched through its web interface [6]. Calculated NMR chemical shift values were obtained from Advanced Chemistry Development, Inc. (ACD/Labs, Toronto, Canada) CNMR Predictor and DB software, version 2020.1.0 [22], as produced for chemical shift checking. The Python computer language interpreter, version 3.7.1, was supplied by Anaconda [23]. Structure file transformations were carried out using the RDKit cheminformatics library [24] and locally developed Python scripts publicly available from the CNMR_Predict directory of the KnapsackSearch Github project [25].

3. Results and Discussion

The creation of a taxonomy focused compound library with predicted ^{13}C NMR chemical shift values included is illustrated here for turnip, or *Brassica rapa* subsp. *rapa* L. Scheme 1 helps to follow the successive steps of the creation process. The related files are available from the Turnip subdirectory of directory CNMR_Predict [25]. Submitting species name *Brassica rapa* as keyword (thus including subspecies other than turnip) for a simple search in the LOTUS DB resulted in 121 hits. The search result was downloaded as file *lotus_simple_search_result.sdf* in V3000 SDF format [26] and stored in a local computer directory, in which all files related to the turnip project were stored. A new ACD/Labs DB file, *turnip.NMRUDB*, was created and filled with data from *lotus_simple_search_result.sdf*. Exporting this database in SDF format as file *turnip.sdf* converted it to the V2000 SDF format. This conversion (Scheme 1, step *a*) is useful if an SDF file reader that does not decode properly the V3000 format is used for structure viewing.



Scheme 1. Database creation using CNMR_Predict: (*a*) initial creation of a multistructure SDF file in V2000 format from a downloaded LOTUS search result; (*b*) SDF file normalization: removal of duplicated structures, tautomer correction, and valence data adjustment for electrically charged atoms; (*c*) creation of a DB with fake chemical shift data; (*d*) incorporation in the DB of the predicted chemical shift values; (*e*) transfer of the predicted chemical shift values for future dereplication studies.

The *turnip.sdf* file then underwent three preparatory operations (Scheme 1, step *b*) before it could be supplemented with chemical shift values. The *turnip.sdf* file may have contained identical structures, apparently because different InChI [27] character strings in LOTUS may have resulted in the production of structures in *lotus_simple_search_result.sdf* that were in turn recoded as identical InChI strings by the RDKit library. A python script, *uniqInChI.py*, retains only a single occurrence of duplicated structures according to InChI equality and was applied to file *turnip.sdf*, in which only one compound out of 121 was 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. As 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 iminol substructures into their amide tautomer, as achieved by the *tautomer.py* script applied to file *turnip.sdf*. It should be noticed that the systematic nomenclature [28] of iminol-containing compounds in LOTUS is determined as if they were really iminols and not amides, resulting, for example, in the difficult identification of peptidic bonds in peptides.

Script *tautomer.py* relies on RDKit to write SDF files and makes use of reaction SMARTS [29]. Electrically charged atoms in structures written by RDKit include a nondefault specification for the nonstandard valence of such atoms (such as four for the nitrogen atom in an ammonium group), in accordance with SDF specification. Such a structure description is not properly interpreted by the ACD/Labs software, thus precluding the prediction of chemical shifts. The *rdcharge.py* script resets the valence data piece to the default, nonblocking value and was applied to *turnip.sdf*.

The first step toward the automatic calculation of ^{13}C NMR chemical shifts (Scheme 1, step *c*) was 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 (Scheme 1, step *d*), a new ACD/Labs DB file, *fake_acd_turnip.NMRUDB*, was created and filled by importation from the file *fake_acd_turnip.sdf*. All carbon atoms appeared 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 displayed a message that warned that the calculated chemical shifts would not be exported. This is simultaneously true and false. This is true because the calculated values cannot be used for a structure search according to the 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 last step toward a DB file usable for dereplication (Scheme 1, step *e*) consists in replacing the 99.99 values under the CNMR_SHIFTS SDF tag that were still present in file *fake_acd_turnip_exported.sdf*, with the calculated chemical shift values it contains, written under the CNMR_CALC_SHIFTS SDF tag. This operation was carried out by the script *CNMR_predict.py* acting on file *fake_acd_turnip_exported.sdf* to produce file *true_acd_turnip.sdf*. A new ACD/Labs DB file, *lotus_turnip.NMRUDB* was finally created and filled with compounds from file *true_acd_turnip.sdf*. The DB *lotus_turnip.NMRUDB* was then ready for compound identification according to ^{13}C NMR chemical shift values using the compound search tool included in the ACD/Labs software. The file *true_acd_turnip.sdf* also contains SDF tags that make dereplication possible by the MixONat software. The script *ACD_to_DerepCrude.py* formatted the predicted chemical shifts for its use with the DerepCrude software. Both DerepCrude [12] and MixONat [13] 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 [10,30].

The values that were calculated for the chemical shift checking of an entire DB file, written under the CNMR_CALC_SHIFTS SDF tag in file *fake_acd_turnip_exported.sdf*, were not exactly those produced by the ACD/Labs CNMR Predictor when run in a compound by compound procedure, but the origin of the difference is difficult to track as no information is available on the details of the underlying algorithms.

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 CNMR_Predict

project files so that the actions to perform sequentially can be easily accomplished. Figure 1 illustrates the content of this template file. CNMR_Predict is a follow up of the Knapsack-Search project that made use of nmrshiftdb2 for the prediction of ^{13}C NMR chemical shift values [31]. 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.

```
    A   New DB    ***.NMRUDB
    C   Import    lotus_simple_search_result.sdf
    D   Export    ***.sdf
        Close
python -m uniqInChI ***.sdf
python -m tautomer ***.sdf
python -m rdcharge ***.sdf
python -m fakefakeACD ***.sdf
    A   New DB    fake_acd_***.NMRUDB
    C   Import    fake_acd_***.sdf
    D   Tools --> Check Chemical Shifts
        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 a name, like "turnip" in the present example.

Creating a file such as *lotus_turnip.NMRUDB* with the ACD/Labs software from initial file *lotus_simple_search_result.sdf* without CNMR_Predict would require a tedious compound by compound operation, lasting about one minute per structure unless some presently undisclosed script is available for calculation process automation [32]. The prediction involving CNMR_Predict lasts less than one second per structure, making it easy to use for the creation of taxonomically focused collections of natural products. For example, species *Brassica rapa* is related to family Brassicaceae and searching for this taxon in LOTUS results in 2271 hits. 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 performs the repetitive work. The choice of the appropriate taxon type (order, family, genus, species, etc.) is left to the user and can be adapted according to the size of the chemical space to investigate. This taxonomy focused approach is more flexible than the one consisting in precalculating the chemical shifts for all the compounds present in a snapshot of a database such as LOTUS since its content may be steadily updated.

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 [33]. The library of the compounds from *Brassica rapa*, a Brassicaceae species, reported by LOTUS contain such erroneous or incompletely defined structures that would be also found in general purpose databases such as that of the Chemical Abstract Service [34].

4. Conclusions

The CNMR_Predict 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.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Beutler, J.A.; Alvarado, A.B.; Schaufelberger, D.E.; Andrews, P.; McCloud, T.G. Dereplication of phorbol bioactives: *Lyngbya majuscula* and *Croton cuneatus*. *J. Nat. Prod.* **1990**, *53*, 867–874. [CrossRef]
2. Hubert, J.; Nuzillard, J.-M.; Renault, J.-H. Dereplication strategies in natural product research: How many tools and methodologies behind the same concept? *Phytochem. Rev.* **2017**, *16*, 55–95. [CrossRef]
3. Sorokina, M.; Steinbeck, C. Review on natural products databases: Where to find data in 2020. *J. Cheminform.* **2020**, *12*, 20. [CrossRef] [PubMed]
4. COCONUT: Natural Products Online. Available online: Coconut.naturalproducts.net (accessed on 5 June 2021).
5. Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M.A.; Steinbeck, C. COCONUT online: Collection of Open Natural Products database. *J. Cheminform.* **2021**, *13*, 2. [CrossRef] [PubMed]
6. LOTUS: Natural Products Online. Available online: Lotus.naturalproducts.net (accessed on 5 June 2021).
7. Rutz, A.; Sorokina, M.; Galgonek, J.; Mietchen, D.; Willighagen, E.; Graham, J.; Stephan, R.; Page, R.; Vondrášek, J.; Steinbeck, C.; et al. Open Natural Products Research: Curation and Dissemination of Biological Occurrences of Chemical Structures through Wikidata. *bioRxiv* **2021**. [CrossRef]
8. Picot-Allain, C.; Mahomoodally, M.F.; Ak, G.; Zengin, G. Conventional versus green extraction techniques—A comparative perspective. *Curr. Opin. Food Sci.* **2021**, *40*, 144–156. [CrossRef]
9. Aron, A.T.; Gentry, E.C.; McPhail, K.L.; Nothias, L.-F.; Nothias-Esposito, M.; Bouslimani, A.; Petras, D.; Gauglitz, J.M.; Sikora, N.; Vargas, F.; et al. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nat. Protoc.* **2020**, *15*, 1954–1991. [CrossRef] [PubMed]
10. Hubert, J.; Nuzillard, J.-M.; Purson, S.; Hamzaoui, M.; Borie, N.; Reynaud, R.; Renault, J.-H. Identification of Natural Metabolites in Mixture: A Pattern Recognition Strategy Based on ^{13}C NMR. *Anal. Chem.* **2014**, *86*, 2955–2962. [CrossRef]
11. Nuzillard, J.-M.; Lameiras, P. Tailoring the nuclear Overhauser effect for the study of small and medium-sized molecules by solvent viscosity manipulation. *Prog. Nucl. Magn. Reson. Spectrosc.* **2021**, *123*, 1–50. [CrossRef]
12. Bakiri, A.; Hubert, J.; Reynaud, R.; Lanthony, S.; Harakat, D.; Renault, J.-H.; Nuzillard, J.-H. Computer-Aided ^{13}C NMR Chemical Profiling of Crude Natural Extracts without Fractionation. *J. Nat. Prod.* **2017**, *80*, 1387–1396. [CrossRef]
13. Bruguière, A.; Derbré, S.; Dietsch, J.; Leguy, J.; Rahier, V.; Pottier, Q.; Bréard, D.; Suor-Cherer, S.; Viault, G.; Le Ray, A.-M.; et al. MixONat, a Software for the Dereplication of Mixtures Based on ^{13}C NMR Spectroscopy. *Anal. Chem.* **2020**, *92*, 8793–8801. [CrossRef]
14. Lauro, G.; Bifulco, G. Elucidating the Relative and Absolute Configuration of Organic Compounds by Quantum Mechanical Approaches. *Eur. J. Org. Chem.* **2020**, 3929–3941. [CrossRef]
15. Joseph-Nathan, P.; del Rio, R.E. Vibrational Circular Dichroism Absolute Configuration of Natural Products From 2015 to 2019. *Nat. Prod. Commun.* **2021**, *16*, 1–30. [CrossRef]
16. Wenzel, T.J. Strategies for using NMR spectroscopy to determine absolute configuration. *Tetrahedron Asymmetry* **2017**, *28*, 1212–1219. [CrossRef]
17. Robien, W. The Advantage of Automatic Peer-Reviewing of ^{13}C-NMR Reference Data Using the CSEARCH-Protocol. *Molecules* **2021**, *26*, 3413. [CrossRef]
18. Bruguière, A.; Derbré, S.; Coste, C.; Le Bot, M.; Siegler, B.; Leong, S.T.; Sulaiman, S.N.; Awang, K.; Richomme, P. ^{13}C-NMR dereplication of Garcinia extracts: Predicted chemical shifts as reliable databases. *Fitoterapia* **2018**, *131*, 59–64. [CrossRef] [PubMed]
19. Lianza, M.; Leroy, R.; Machado Rodrigues, C.; Borie, N.; Sayagh, C.; Remy, S.; Kuhn, S.; Renault, J.-H.; Nuzillard, J.-H. The Three Pillars of Natural Product Dereplication. Alkaloids from the Bulbs of *Urceolina peruviana* (C. Presl) J.F. Macbr. as a Preliminary Test Case. *Molecules* **2021**, *26*, 637. [CrossRef] [PubMed]
20. KNApSAcK Family Top Page. Available online: Knapsackfamily.com (accessed on 5 June 2021).
21. Beniddir, M.A.; Kang, K.B.; Genta-Jouve, G.; Huber, F.; Rogers, S.; van der Hooft, J.J.J. Advances in decomposing complex metabolite mixtures using substructure- and network-based computational metabolomics approaches. *Nat. Prod. Rep.* **2021**. [CrossRef]
22. Chemistry Software for Analytical and Chemical Knowledge Management. Available online: Acdlabs.com (accessed on 5 June 2021).
23. Anaconda | The World's Most Popular Data Science Platform. Available online: Anaconda.com (accessed on 5 June 2021).
24. RDKit. Available online: Rdkit.org (accessed on 5 June 2021).
25. nuzillard/KnapsackSearch: Automated Data Search in the KNApSAcK Database. Available online: Github.com/nuzillard/KnapsackSearch (accessed on 5 June 2021).
26. ctfiles.book—ctfile.pdf. Available online: Daylight.com/meetings/mug05/Kappler/ctfile.pdf (accessed on 5 June 2021).

27. Heller, S.R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminform.* **2015**, *7*, 23. [CrossRef] [PubMed]
28. Blue Book-IUPAC | International Union of Pure and Applied Chemistry. Available online: Iupac.org/what-we-do/books/bluebook (accessed on 5 June 2021).
29. Daylight Theory: SMARTS-A Language for Describing Molecular Patterns. Available online: Daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed on 5 June 2021).
30. NatExplore. Available online: Nat-explore.com (accessed on 5 June 2021).
31. Steinbeck, C.; Kuhn, S. NMRShiftDB—Compound identification and structure elucidation support through a free community-built web database. *Phytochemistry* **2004**, *65*, 2711–2717. [CrossRef] [PubMed]
32. Richomme, P. *Personal Communication*; University of Angers: Angers, France, 2020.
33. Ibrahim, N.; Allart-Simon, I.; De Nicola, G.R.; Iori, R.; Renault, J.-H.; Rollin, P.; Nuzillard, J.-M. Advanced NMR-Based Structural Investigation of Glucosinolates and Desulfoglucosinolates. *J. Nat. Prod.* **2018**, *81*, 323–334. [CrossRef] [PubMed]
34. Empowering Innovation & Scientific Discoveries. Available online: Cas.org (accessed on 5 June 2021).