Using Predicted ^{13}C NMR Spectra with Open Resources for Structure Dereplication

Dimitris Argyropoulos[1], Sergey Golotvin[1], Rostislav Pol[1], Arvin Moser[1], Jessica Litman[1], Nico Ortlieb[2], Steffen Breinlinger[2], Tomasz Chilczuk[2] and Timo H. J. Niedermeyer[2]

1. **Advanced Chemistry Development, Inc.**
Toronto, ON, Canada
www.acdlabs.com
2. **Institute of Pharmacy, RG Pharmacognosy, Martin-Luther-University Halle-Wittenberg**
Halle (Saale), Germany

Introduction

Over the past two decades, market pressure has led to increased demands for development of New Molecular Entities (NME's).[1] In response, the pharmaceutical industry has attempted to accelerate this by implementing more efficient, higher volume techniques into development procedures. These include, high-throughput screening, parallel synthesis, and absorption, distribution, metabolism, and excretion toxicology (ADMET) predictions. Natural product discovery programs reveal chemical diversity that can complement high-throughput screening efforts. However, this is only worthwhile if the active components in natural product mixtures can be reliably separated and quickly identified. The practice of screening active compounds early in the development process for recognizing and eliminating known compounds is called dereplication. This enables scientists to focus on testing truly 'unknown' compounds.

There are two conditions that must be fulfilled for efficient dereplication:

1. One must be able to easily identify characteristic spectral 'fingerprints' of unknown compounds.
2. One must have access to databases containing spectra of known structures.

NMR and MS spectra are typically used for dereplication. High resolution MS is the simplest and fastest to record, but it lacks the structural information that NMR provides. ^{1}H NMR is a fast and straightforward technique that includes structural information. However, a ^{1}H NMR spectrum is not a reliable fingerprint because of its limited resolution and the fact that measured spectra can be affected by factors like pH, concentration, and solvent effects. The ^{13}C NMR spectrum of a compound, on the other hand, can be considered an effective fingerprint since it is virtually unaffected by the aforementioned conditions. It is also largely magnetic field independent, since there are no couplings that could cause variations in stronger or weaker fields. As a result, it is very easy to predict accurately.

To satisfy the second condition, one can consider using databases of real spectra or predicted spectra. Databases of real spectra usually contain a limited number of structures, and their spectra may not be ideal. On the other hand, there are several "open" databases with millions of chemical structures that could be used to predict ^{13}C spectra, an example is PubChem.[2] The benefits of using predicted spectra are that they are magnetic field independent, can be adjusted for solvents, and can be very accurate depending on the algorithms used.[3]

Here we propose an efficient dereplication strategy, which treats the experimental ^{13}C NMR spectra of an unknown as a fingerprint. The 'unknown' is identified by finding a match for its fingerprint in a database of predicted ^{13}C NMR spectra of known compounds. We created this database using the structures from PubChem, a database that allows users to freely download known chemical structures. We explore the possibilities and limitations of using predicted ^{13}C spectra for structures from open databases, describe the workflow, and critically evaluate the usefulness of this technique.

Method

In order to evaluate the proposed dereplication strategy, one must consider the quality of the predicted ^{13}C spectra. To ensure that the spectra of the PubChem structures are predicted as accurately as possible, we used ACD/Labs NMR Predictors – the industry standard for NMR prediction.[3,4]

The following parameters must be defined to ensure that spectral matches are not falsely excluded from the search results:

1. The mean maximum difference between experimental and theoretical chemical shifts. This is usually set as ≤ 2 ppm.
2. The number of known peaks in the experimental spectrum that are not visible due to a low signal to noise ratio (i.e. number of missing peaks). This value can be set to 0 or 1 for good quality spectra or higher otherwise.
3. The number of peaks resulting from impurities (number of extra peaks). Can be low if only the appropriate peaks are picked.
4. Optional – one can filter by molecular formula to expedite the search time. For example, this can reduce a ten minute search to 30 seconds.

Experimental and Results

To illustrate the benefits and limitations of this dereplication strategy, we show 4 examples of results obtained while attempting to identify an unknown substance.

Case 1 | Identification of a Known Compound

Ergosta-7,22-diene-3-one was isolated from mature fruiting bodies of the basidiomycete *G. pfeifferi*, which were collected in the vicinity of Greifswald (Germany). They were air-dried and powdered, then extracted and separated by chromatography. High-resolution mass spectrometry (HR MS) indicated a molecular formula of $C_{28}H_{44}O$. The 1D-^{13}C spectrum had 28 peaks, and using the strategy outlined above, the compound was identified within ~30 seconds as 7,22-Ergostadienone (PubChem ID6436804), the search results also included all its stereoisomers.[5]

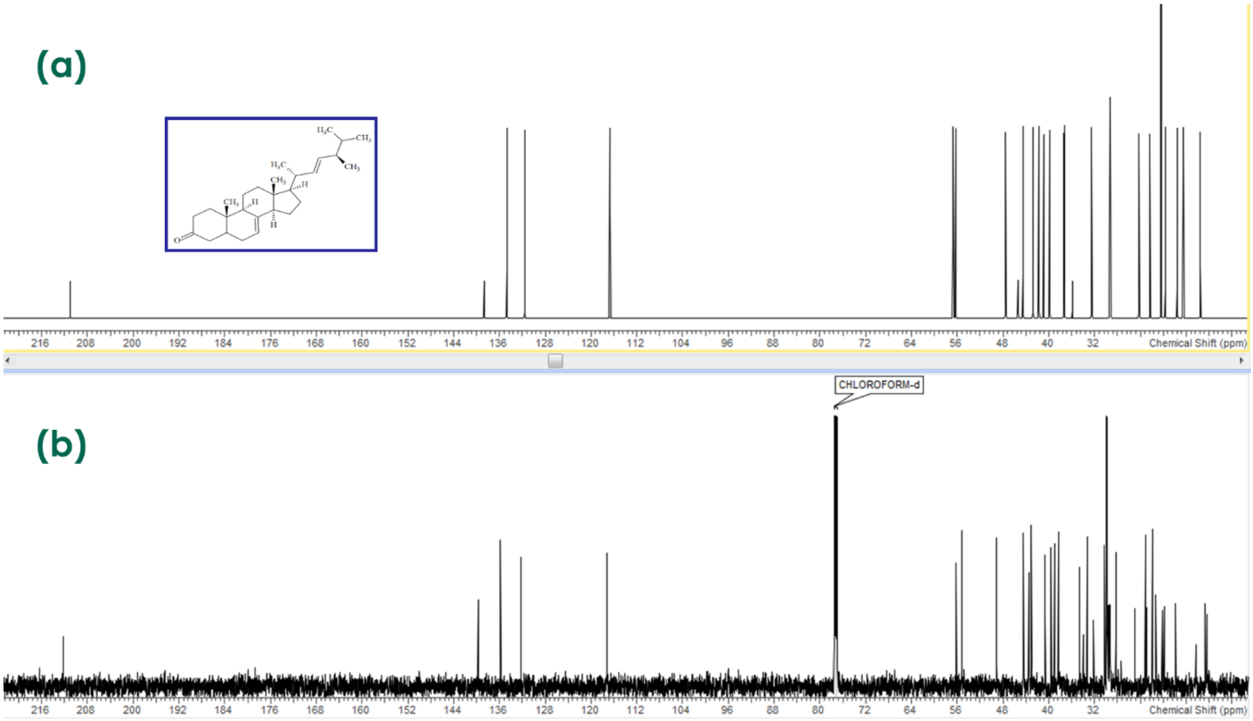

Figure 1. (a) Predicted ^{13}C spectrum of 7,22-Ergostadenone (PubChem ID6436804) (b) Experimental ^{13}C spectrum of a natural compound extracted from the mature fruiting bodies of the basidiomycete G. pfeifferi.

Case 2 | Identification of an Unknown Compound

Dihydroxanthoxidin was isolated from a *Streptomyces* sp. strain, *Streptomyces* sp. AcE210, which was isolated from root nodules of a black alder tree (*Alnus glutinosa*). The strain was cultivated and extracted, and the produced compounds were isolated via chromatographic methods. HR MS indicated a molecular formula of $C_{11}H_{18}O_5$, and the 1D-^{13}C spectrum contained 11 peaks. For dereplication, the outlined strategy with standard settings resulted in zero hits, therefore a complete structure elucidation was performed. This revealed the structure of dihydroxantoxidin, which does not exist in PubChem [6]. The related ketone xanthocidin exists in PubChem (ID 206670). This shows that small structural differences are enough to exclude false positives. Despite the similarity of the structures, the NMR spectra for dihydroxantoxidin and xanthocidin possess significant differences. Repeating the search with more relaxed tolerances for missing or extra peaks, yields a few hundred hits. Xanthocidin was amongst them, but it was not well ranked by mean experimental vs. predicted chemical shift deviation (i.e. differences >2 ppm)

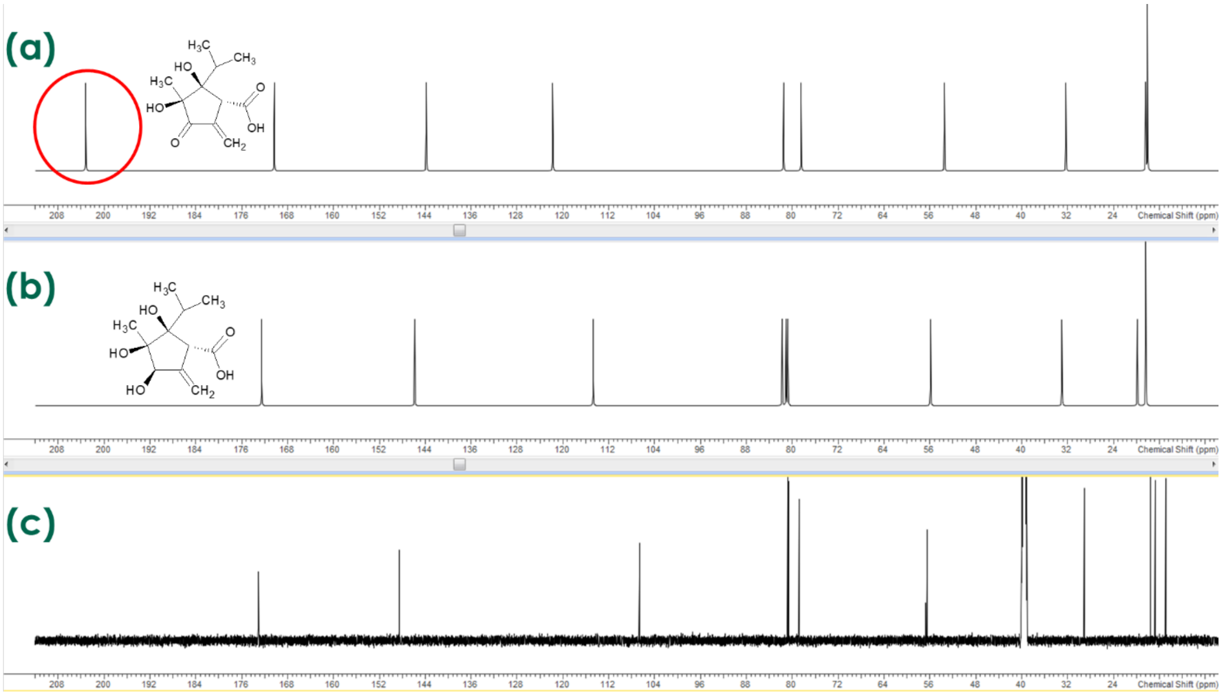

Figure 2. (a) Neural Network predicted database spectrum of xanthocidine, (b) Neural network predicted database spectrum of dihydroxanthocidine and (c) experimental ^{13}C NMR spectrum of dihydroxanthocidine. Apart from the difference of the carbonyl peak at 210 ppm (a) and the alcohol group at 80 ppm (b) region there are no other significant differences between the spectra of the two compounds.

Case 3 | Verifying an Unknown using only a 2D-^{13}C NMR Spectrum

In natural product research it is not uncommon to lack sufficient isolated sample volume to record a 1D-^{13}C spectrum. This was the case with the compound used to record Figure 3, which was isolated from a cyanobacterium, *Nostoc* sp. The compound was extracted and then further purified by chromatography. 1D-^{1}H, 2D HSQC and 2D HMBC spectra were all recorded. The dereplication strategy outlined above can still be employed without a 1D-^{13}C spectrum, as the table of spectral data containing all the ^{13}C resonances observed in the 2D spectra can be used instead. After <1 minute of searching no definite matches were found in the database, i.e., the mean deviation between experimental and predicted chemical shifts was not lower than 2 ppm for any known structures. This confirmed that the structure was indeed novel.[7] This illustrates the appreciable time that can be saved using this technique since a comprehensive manual search would have taken a few days/weeks.

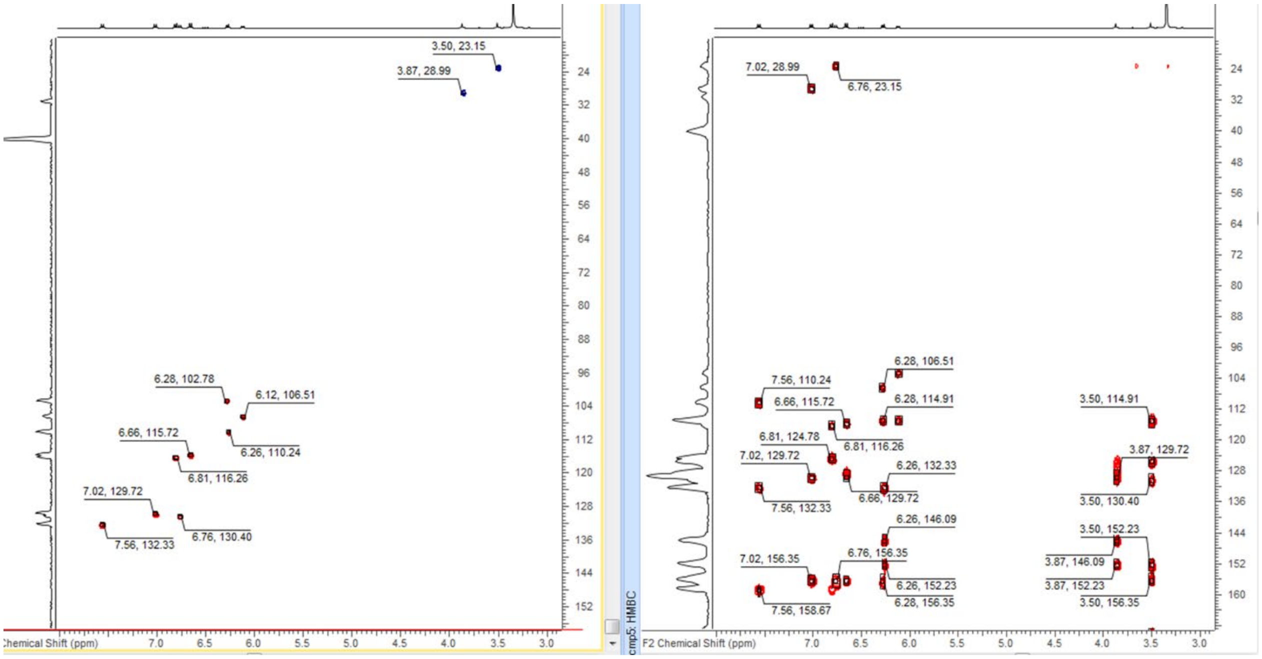

Figure 3. HSQC (left) and HMBC (right) spectra of the novel compound recorded over a period of about 21 and 18 hours respectively. It would have been impractical to record a 1D-^{13}C.

Case 4 | Identification of a Famously Misidentified Compound

Baulamycin-A is famously known for having its 3D structure and stereochemistry misidentified multiple times. Only recently, Butts *et al* (2017) identified its true structure using a collection of spectroscopic, computational (i.e. density functional theory), and synthetic methods.[8]

Searching the predicted spectra for PubChem structures produces two results: IDs 100951617 and 74223134. Neither is named Baulamycin. A search of the PubChem database for Baulamycin-A results in the entry shown in Figure 4(a), which is an incorrect structure. The correct structure has a 3-methyl butyl group rather than a 2-methyl butyl as shown in PubChem.

Even if one is able to identify a structure as known, discretion is required since the dereplication method is not immune to mislabeled structures in open source databases. As a result, one must critically consider the structural matches, particularly when multiple structural results are given for a single spectrum.

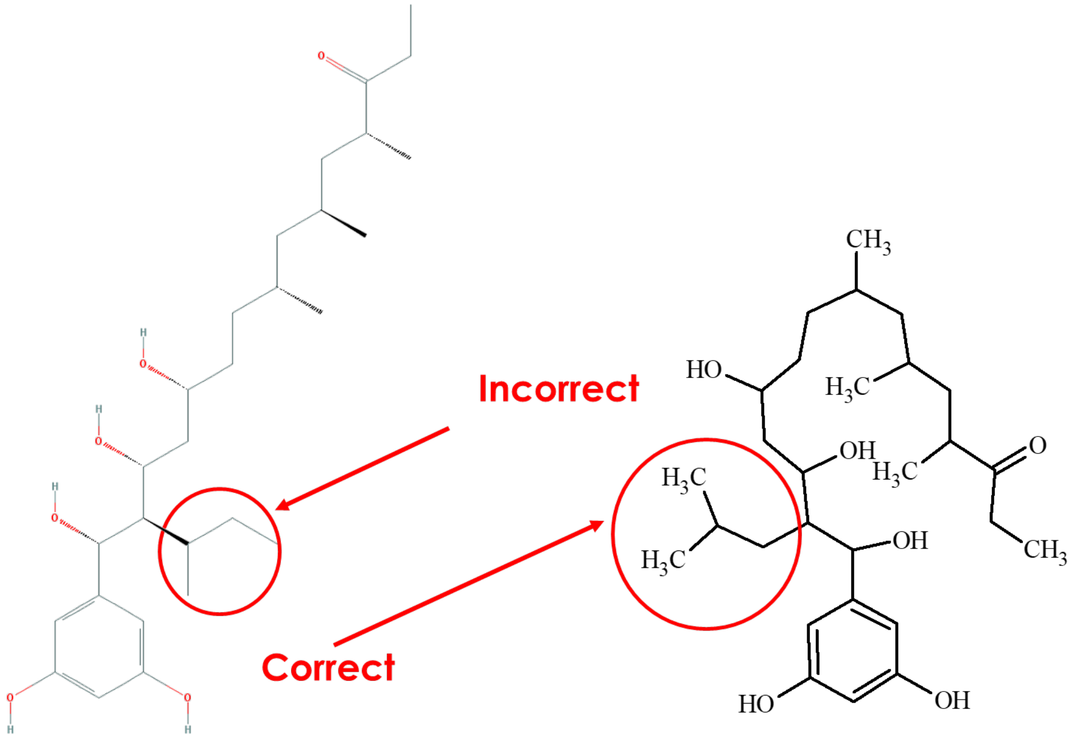

Figure 4. The correct and incorrect structures of Baulamycin-A. The incorrect structure is shown in PubChem with a 2-methyl butyl group. The correct structure has a 3-methyl butyl group.

Discussion and Conclusions

This dereplication strategy using ^{13}C spectra is a very powerful method since it is able to significantly reduce the time spent determining if an 'unknown' has previously been identified. One can see that databases of accurately predicted ^{13}C spectra are a very strong resource for dereplication, and that PubChem is an invaluable source of reported structures. However, there is always a possibility of error, so some caution is needed when interpreting search results.

References

1. Earm, K., Earm, Y. E., Integrative Medicine Research, **2014**, 3, 211-216.

2. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH. PubChem Substance and Compound databases. Nucleic Acids Res. **2016** Jan 4; 44(D1):D1202-13. Epub **2015** Sep 22 [PubMed PMID: 26400175] doi: 10.1093/nar/gkv951.

3. Data presented by Burkhard Kirste, FU Berlin, 38th FGNMR Meeting, Sept. **2016**, Dusseldorf

4. ACD/Labs, "ACD/NMR Predictors," **2017**. [Online]. Available: www.acdlabs.com/nmrpredictors

5. T.H.J. Niedermeyer, U. Lindequist, R. Mentel, D. Gördes, E. Schmidt, K. Thurow, M. Lalk. Antiviral Terpenoid Constituents of Ganoderma pfeifferi. Journal of Natural Products, 68(12): 1728-1731, 2005.

6. N. Ortlieb, K. Bretzel, A. Kulik, J. Haas, S. Lüdeke, N. Keilhofer, S.D. Schrey, H. Gross, T.H.J. Niedermeyer. Xanthocidin Derivatives from the Endophytic Streptomyces sp. AcE210 Provide Insight into the Xanthocidin Biosynthesis. 19(23): 2472-2480, 2018.

7. Manuscript in preparation

8. Jingjing Wu, Paula Lorenzo, Siying Zhong, Muhammad Ali, Craig P. Butts, Eddie L. Myers & Varinder K. Aggarwal, Nature, **2017** *547*, 436–440.