Searching Libraries of Known Structures for Dereplication: Benefits and Requirements

Nadia Laschuk, Dimitris Argyropoulos, Rostislav Pol, and Sergey Golotvin
Advanced Chemistry Development, Inc. (ACD/Labs), Toronto, ON, Canada

Introduction

Dereplication is an important step in natural product research as well as in competitive, counterfeit, and reaction product analysis. It aims to answer the question

"Has anyone seen this structure before?"

Dereplication can be done using LC-MS (retention time and molecular ion mass) or NMR data, amongst others. The use of ^{13}C NMR spectral information is generally preferred[1] as it provides a clear fingerprint of the compound's carbon skeleton. We have previously explored the benefits of using databases with predicted spectra of compounds. Currently, both commercial[2] and freely available[3-5] options exist. Here, we explore the capabilities and requirements of such systems.

Results and Discussion

We selected 56 compounds from the Aldrich library of FT-NMR spectra, with a molecular weight (MW) range of 100-800. We then searched the database of Known Structures available with ACD/NMR Workbook, consisting of >100 million predicted spectra for structures from PubChem. Using the observed experimental peak frequencies and the MW, the correct structure was identified as the top hit in the majority of cases (Figure 2), indicating the validity and potential of the method.

For the cases where the correct structure was not ranked first, we experimented with adjusting the number of extra or missing peaks allowed in the experimental data input. (Figure 1) In almost all the cases, we were able to find optimal conditions that gave the correct result.

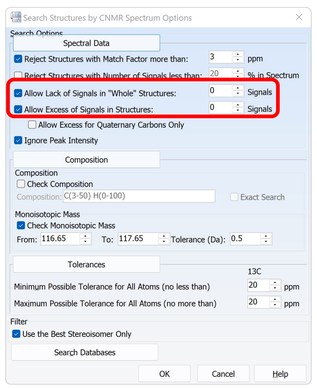

Figure 1: The search options dialog box, highlighting the fields to enter the values for the extra and missing peaks in the experimental spectrum.

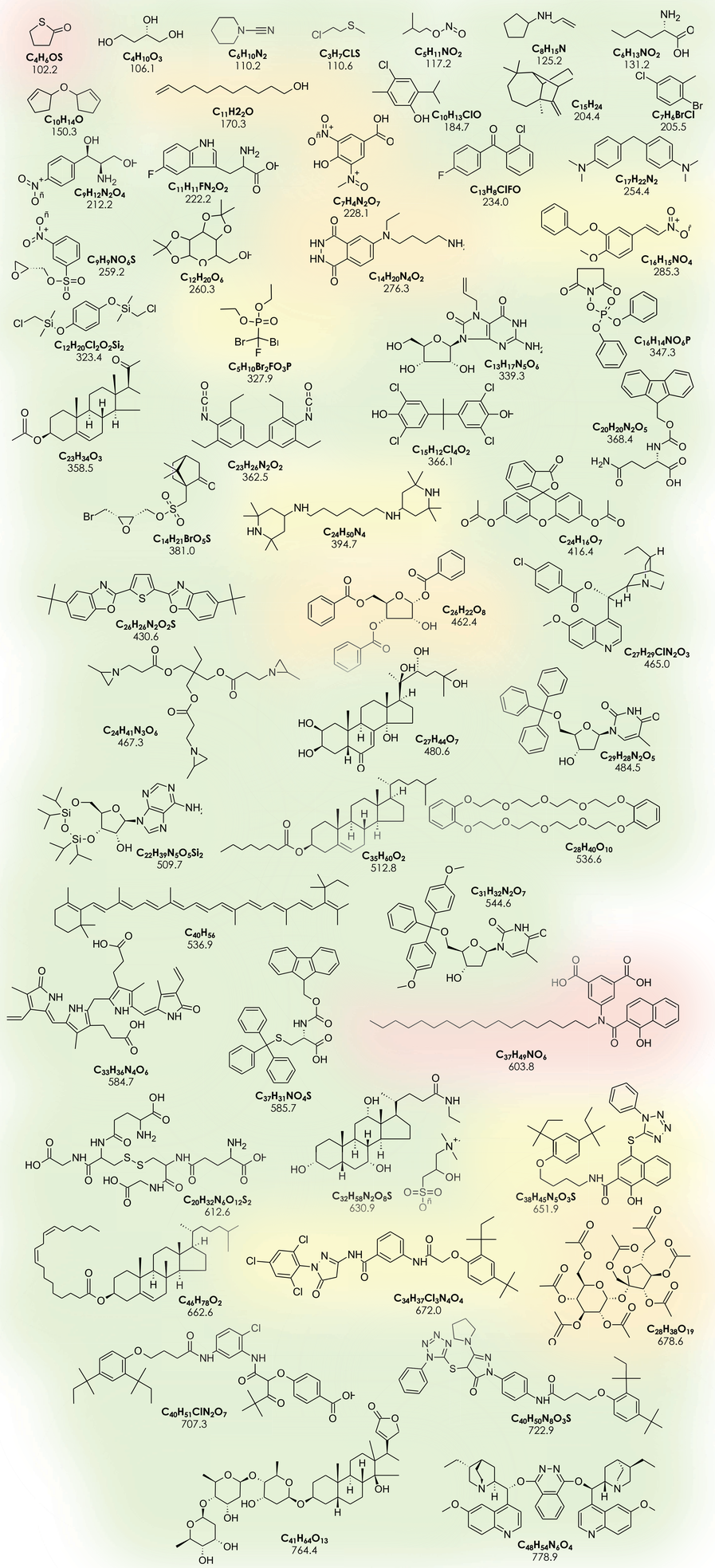

Figure 2: The chemical structures used in this study. Those highlighted in green were ranked first in the search, those in yellow were second, those in orange were ranked in positions 3 to 9, and those in red were ranked 10th or higher.

We also saw that as the MW increases, the uncertainty and the number of missing or extra peaks required increases as well. (Figure 3) However, the correct result can always be identified within a few seconds.

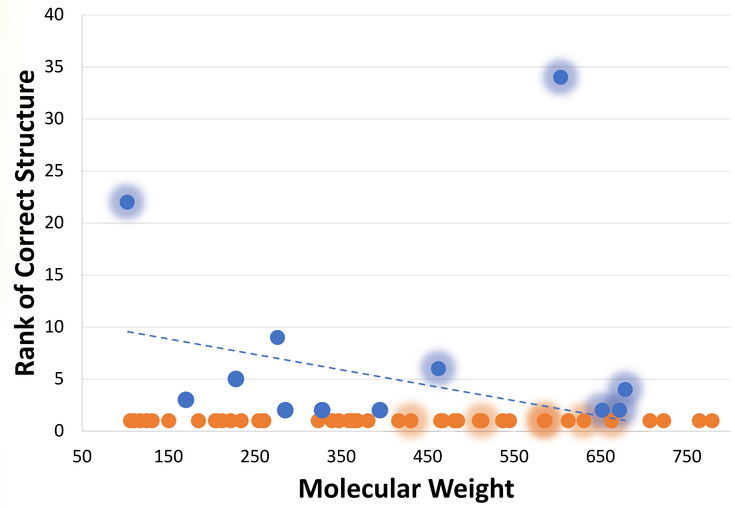

Figure 3: Rank of the correct structure in the search results vs. MW for all the compounds. Orange dots represent hits ranked 1 while blue dots represent ranking more than 1. Dots with a shadow required search parameters to allow one or more missing or extra peaks.

Finally, we explored the search options with respect to the inclusion/exclusion of the MW information. We saw that the MW information is essential, as it provides a very clear starting point and deals effectively with symmetric compounds.

Conclusion

Searching libraries of predicted ^{13}C NMR spectra is a highly efficient dereplication method that gives the correct result in seconds when the search parameters are sufficiently optimized. We also found that the inclusion of the MW is essential for the search, and increasingly relaxed search options are required as the MW increases.

References

1. R.B. Williams, M. O'Neill-Johnson, A.J. Williams, P. Wheeler, R. Pol, A. Moser. (**2015**). *Org. Biomol. Chem., 13,* 9957–9962.
2. D. Argyropoulos, S. Golotvin, R. Pol, A. Moser, N. Ortel, S. Breinlinger, T. Chilzuk and T. Niedermeyer, "Efficient Dereplication of Natural Products Using Predicted 13C Spectra" presented at 60th ENC, Poster 004, 2019.
3. H. Kalchhauser, W. Robien. (**1985**). *Chem. Inf. Comput. Sci., 25*(2), 103–108.
4. R. Reher, H. W. Kim, C. Zhang et al. (**2020**). *J. Am. Chem. Soc., 142*(9), 4114–4120 .
5. Z. Yang, J. Song, M. Yang, L. Yao, J. Zhang, H. Shi, X. Ji, Y. Deng and X. Wang. (**2021**). *Anal. Chem., 93*(50), 16947–16955.

@ACDLabs info@acdlabs.com www.acdlabs.com US/Canada 1 800 304 3988 Europe +44 (0) 1344 668 030

Visit the ACD/Labs Hospitality Suite in Poinciana B