Dereplication of Natural Products by NMR: A Three-Stage Approach

ACD/Structure Elucidator Suite and ACD/Labs' Content Databases

Arvin Moser, Patrick Wheeler, Steve Hayward
Advanced Chemistry Development, Inc.
Toronto, ON, Canada
www.acdlabs.com

Introduction

For many years, the main source of marketed drug therapies was natural products, or their semi-synthetic derivatives. Over the past two decades, pharmaceutical firms have faced increasing market pressure to accelerate the discovery and development of New Molecular Entities (NMEs). Consequently, firms have implemented a variety of strategies in response to this pressure (*e.g.,* High-Throughput Screening, Parallel Synthesis, and Predictive ADMET). The perception that natural product isolation/characterization is inefficient has caused many firms to abandon their natural product discovery efforts.

However, missing in this new paradigm is the broad chemical diversity that natural product scaffolds provide. Some organizations and academic groups are now bringing natural product discovery programs back as a complement to their other discovery efforts. In order to be successful in natural product-based drug discovery, the capability to quickly and reliably separate and identify the active components in natural products in mixtures—identified through bio-assay and/or mass spectrometry guided fractionation—is a critical need. Dereplication refers to the process of screening active compounds early in the development process to recognize and eliminate those compounds that have been studied in the past, thereby proactively decreasing the number of structures that will need to be fully elucidated and minimizing the amount of time spent on testing.

This application note will focus on a complete system and workflow for dereplication by NMR that takes less than 15 minutes [1] on average (Figure 1) and can help ensure that the time invested in each ensuing elucidation is well spent.

The Dereplication Workflow

In years past, a full complement of spectral data were required to fully characterize the active constituents in component mixtures. Considering the resources required in generating this level of data, and moreover to elucidate the chemical structures within the mixture, a much more efficient process was needed. This multi-stage dereplication process is outlined below and explained in more detail in the section to follow. See the flowchart in Figure 1 for a visual representation of this workflow.

To avoid collecting unnecessary NMR data, a ^{1}H NMR spectrum is all that is required for the first stage. Therefore, precious spectrometer and scientists' time is reserved solely for novel chemical scaffolds.

> **Note** MS dereplication is also available from ACD/Labs. See http://www.acdlabs.com/products/spectrus/workbooks/ms/ms-structure-id/

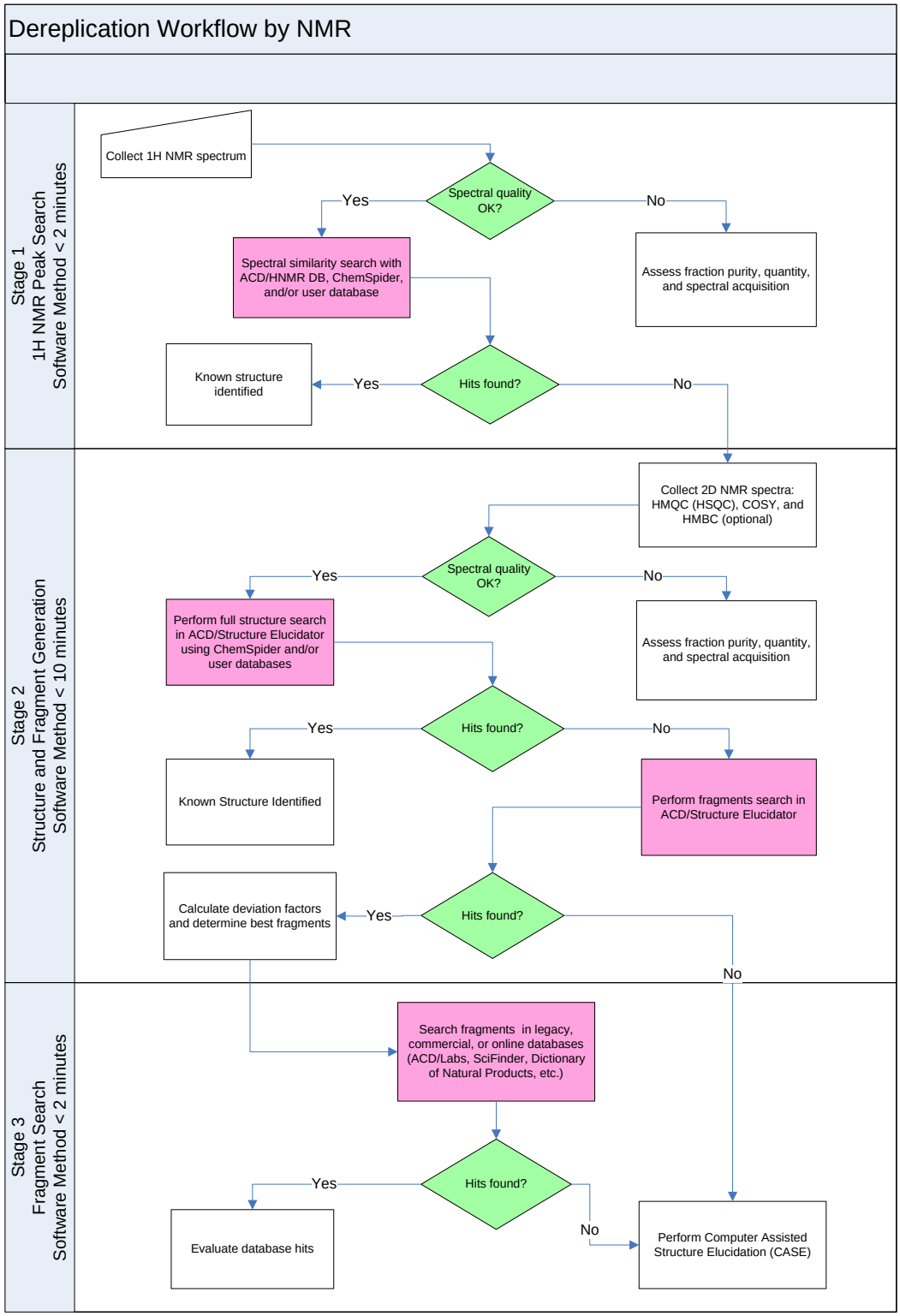

***Figure 1**: An outline of the dereplication process using ACD/Labs' Structure Elucidator and Content databases.*

> **Note** Based on this workflow, a full complement of appropriate spectral data is only required when an isolated structure remains unidentified.

The starting point for this workflow (after spectral acquisition) begins with a ^{1}H NMR spectrum and ACD/HNMR DB [2]. With this data, a chemical shift search can be performed against the 165,000 structures with assigned shifts in the HNMR DB to determine if the NMR data of the unknown is consistent with a structure that is in the database. If this process yields no hits, it is suggested that 2D NMR data be obtained to assist in the next stage of the workflow.

The 2D NMR data can be used in ACD/Structure Elucidator Suite [3] to search its own unique library of molecules. Any resulting hits should be further evaluated by the Chemist. If no hits are generated from the full structure search, the program can then be directed to search its library of structural fragments with approximate chemical shifts. Once relevant fragments are identified, their corresponding match factors can be calculated and the most relevant fragments will be sorted in order to provide easy evaluation for the user.

In the final stage, any unique structural fragments identified by the fragment search can be queried by substructure in several of ACD/Labs' content databases. These databases contain a large proportion of natural product structures and spectral information. These fragments can also be used for substructure searches in online databases such as the Dictionary of Natural Products [4] or SciFinder® [5]. In the rare case that these dereplication methods do not provide satisfactory information to the user, a full-scale Computer-Assisted Structure Elucidation (CASE) can commence.

> **Note** This application note is dedicated to the dereplication aspect of this workflow. Information on structure elucidation can be found on our website.

The following software modules are integral to the dereplication workflow defined here:

- ChemSpider [6] from the Royal Society of Chemistry (a local database of predicted ^{1}H and ^{13}C NMR chemical shifts for *ca*. 22 million chemical structures—a small proportion of these being natural products and estimated to be about 0.2% of the collection [1])
- ACD/HNMR DB [2] (experimental ^{1}H NMR chemical shifts and coupling constants for over 165,000 chemical structures)
- ACD/CNMR DB [2] (experimental ^{13}C NMR chemical shifts and coupling constants for over 165,000 chemical structures)
- ACD/NNMR, FNMR, and PNMR DB [2] (experimental ^{15}N, ^{19}F, and ^{31}P chemical shifts and coupling constants for over 60,000 chemical structures)
- ACD/Structure Elucidator [3] (CASE (Computer-Assisted Structure Elucidation) software package that enables the process of dereplication as a first step by searching its own unique, experimental library of 200,000 molecules and 1.6 million fragments.

Stage 1: Searching for Structures Based on ^{1}H NMR Chemical Shifts

As noted above, ACD/Labs offers several content databases for the process of dereplication. The flexible searching options provided are the key to simplifying and speeding up the process significantly. HNMR DB and CNMR DB (along with NNMR, FNMR, and PNMR DB) are examples of non-spectral content databases that contain millions of chemical shifts, coupling constants, and the corresponding chemical structures. These databases each contain numerous published natural product structures and can be accessed easily through a multitude of searching capabilities that aid specifically in dereplication. Below we describe the first stage of our workflow, and the dereplication of an unknown fraction based on its ^{1}H NMR spectrum using HNMR DB.

ACD/Spectrus Processor [7] in conjunction with HNMR DB offers the ability to search the database by chemical shift values directly from the processing window. By simply picking the peaks of interest in the spectrum, the user can quickly search the database based on the chemical shifts of the peaks highlighted. After a search is executed, the database interface is presented with the hit results. In this example, the peak search generated 7068 hits in the HNMR DB.

At this stage, other queries can be performed to narrow down the search. If for example the molecular formula is known, this information can be used to further refine the search. Figure 2 below, illustrates the results of the molecular formula search in the table view. The software has identified that the published chemical shifts and molecular formulae of four compounds are consistent with the experimental ^{1}H NMR data of the unknown compound. While the coincidence of the chemical shifts and molecular formulae may be good, the onus always falls on the expert to confirm whether a proposed structure is correct or not. In the case where there is any doubt on the user's part, more experiments should be performed on the isolate of interest.

Please keep in mind that there are several other queries that can be used depending on the information that is available (mass, substructure, coupling constant, etc.). In addition, Dereplication using ^{13}C, ^{15}N, ^{19}F, ^{31}P from 1D or 2D NMR data can also be performed using the respective database products and following the same workflow described above. Perhaps the most valuable database to query would be one built on a company's legacy data. This database would ideally include all compounds elucidated within the company in the past and would help avoid multiple elucidations on the same compound.

The next stage of the process will elaborate on what steps would be required if the database search **does not** provide a correct answer.

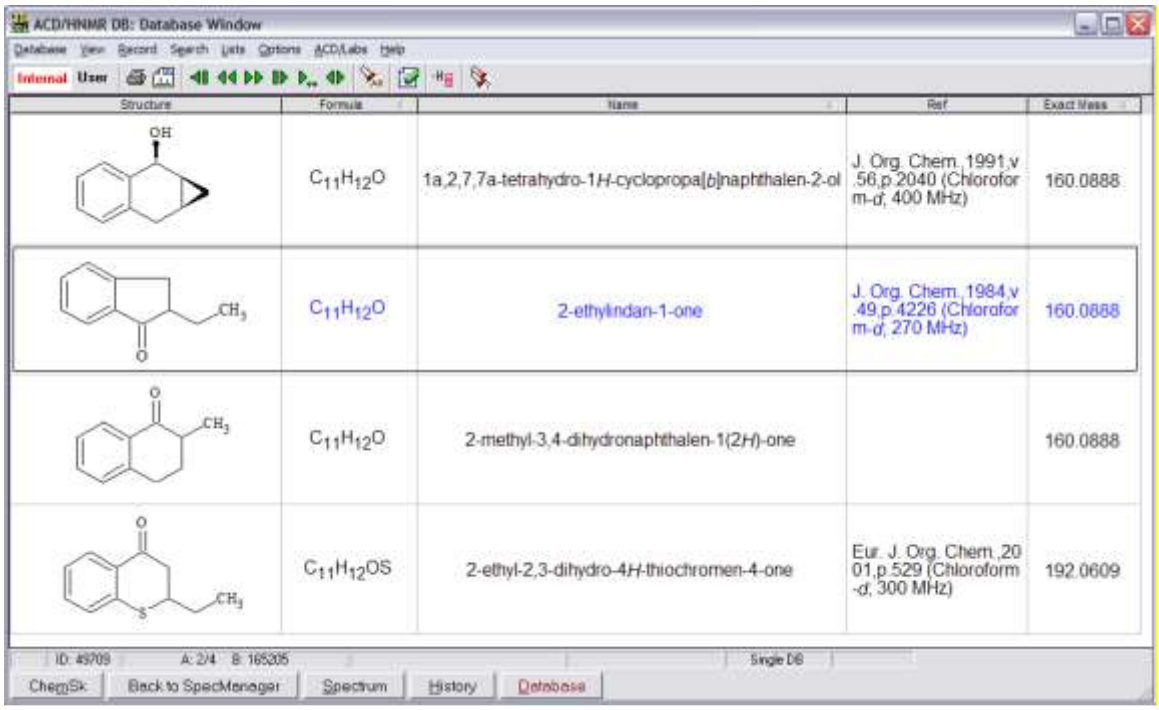

***Figure 2**: The chemical shift search reveals four hits as possible chemical structures of the unknown fraction. In this example the correct structure for the ^{1}H NMR spectrum used was 2-ethylindan-1-one (Compound 2 in the table).*

Stage 2: Generating Structures and Fragments from Spectral Data

In the case where a simple database search based on ^{1}H NMR chemical shifts is not successful, ACD/Structure Elucidator Suite can provide a quick means to identify whether the unknown is in fact a known chemical structure [1,8]. This method is particularly useful because it allows the user to input various types of analytical and chemical data. (1D and 2D NMR data, MS data, IR Data, molecular formula, *etc.*)

Dereplication is a logical step in the elucidation process. As a result, it is performed automatically by the software before any elucidation is considered. By using several types of analytical and chemical data, Structure Elucidator Suite can search its own library of assigned chemical structures and structural fragments, based on the ^{13}C NMR shifts of the sample of interest.

> **Note** At the present time, library searches in Structure Elucidator Suite are only searchable by ^{13}C NMR shifts. Please note that indirectly detected ^{13}C NMR shifts from 2D NMR experiments are sufficient when directly detected ^{13}C shifts are not available (therefore a ^{13}C NMR spectrum need not be obtained).

Figures 4 and 5 illustrate the options and output, respectively, for the dereplication process in Structure Elucidator Suite. The results of the queries provide suggested molecules and/or molecular fragments of the unknown fraction being elucidated.

In this example, we will use a dataset consisting of the following data:

- ^{1}H NMR
- HMBC
- HSQC
- Molecular formula

ACD/Structure Elucidator Suite is built on two structural libraries. The first library contains assigned chemical structures or molecules. The second library contains assigned structural fragments. A search of the compound library resulted in no hits. This means that there are no compounds in the library that match all of the data provided.

Since the program cannot suggest a full molecule in this case, the next step is to search the library of molecular fragments to gain insight regarding what structural fragments may be part of the unknown compound's structure. Figure 3 shows the dialog box and the options available when searching the library of fragments. Note that in order to use information from the 2D NMR data in the search, the appropriate check boxes should be selected.

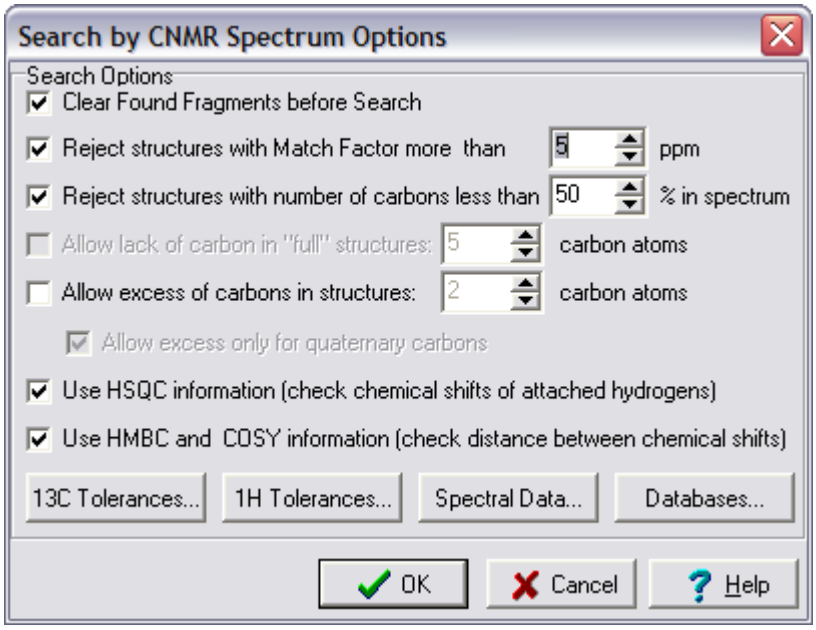

***Figure 3**: Dereplication is performed with ACD/Structure Elucidator Suite by searching the library of fragments based on the spectral data provided. Note that the dialog box allows you to set specific search parameters as well as include information from the 2D NMR data (HMBC and HSQC in this case) to reduce the number of hits.*

A two-minute search of the fragment library in Structure Elucidator Suite suggests three possible molecular fragments based on the data provided (Figure 4). As is often the case, more than one possible match has been generated. The program allows for further filtering of database hits by comparing and reporting the difference between the experimental data and the data from the database hits as a deviation statistic. A general rule of thumb is to reject any proposed fragments with an average shift deviation of greater than 5 ppm. As well, it is important to note that ranking structures with match factors less than 5 ppm should be done with extreme care and the user should evaluate the suggested fragments very carefully. The lower the match, the better the correspondence the fragment has with

the experimental data. Based on this, the software suggests that fragment 1 (ID: 7) is the most consistent with the experimental data provided.

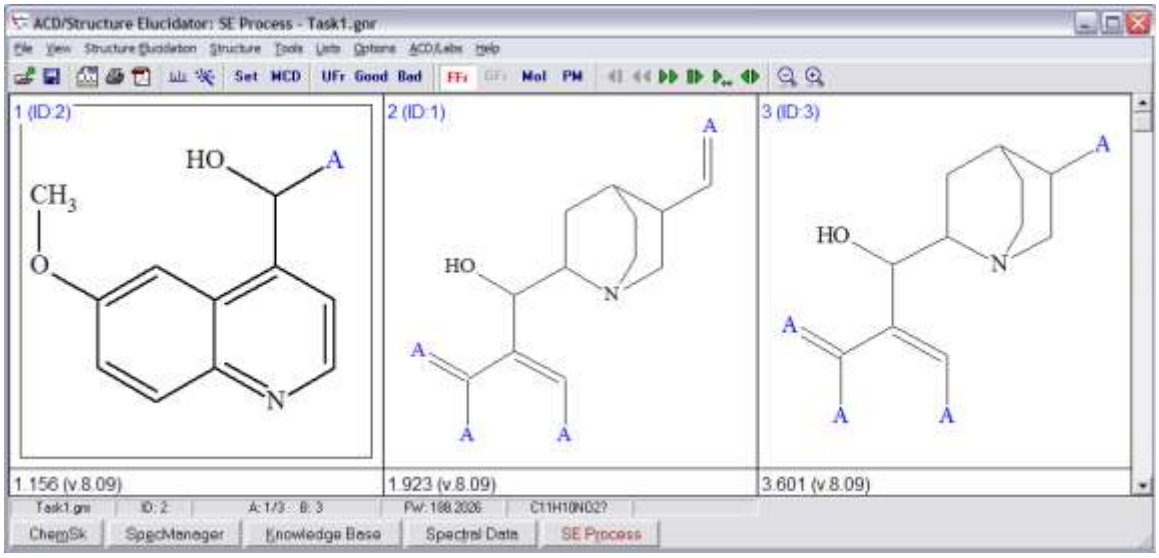

***Figure 4**: Three structural fragments were identified in ACD/Structure Elucidator by searching the library, based on the spectral data provided. Note the Match values for each structure are shown below the corresponding fragment.*

The top ranked fragments resulting from the search can then be used to query structural databases to try and identify the chemical structure of the unknown. The final stage of this process is highlighted in the next section, which will illustrate how this can be done using ACD/Labs' Content Databases.

Stage 3: Use Molecular Fragments to Perform Sub-Structure Searches

Once potential fragments are identified using ACD/Structure Elucidator Suite these fragments can easily be searched in ACD/Labs' Content Databases. The following example will show how to use the top-ranked fragment from the example above in a sub-structure search of ACD/HNMR DB.

In Stage 2, the software suggested 3 possible fragments based on the NMR data provided. In this example we will search the top ranked fragment using a substructure query in ACD/HNMR DB. Searching this fragment produces 40 hits in a matter of seconds. The user can now choose to evaluate the 40 structures to see if they are consistent with the data provided, or they can further narrow the hits by searching another suggested fragment. Doing a second substructure search with the next highest ranked fragment from Structure Elucidator Suite reduces the number of hits to 15, a much more manageable number (alternatively, you can search more than one substructure at a time). Keep in mind that this search can be even further refined through an array of other data search queries. For example, recall that in this problem we had a molecular formula for our unknown. Further refinement of the search with this knowledge results in only 6 hits—a very manageable number. The results are shown in tile

format in Figure 5. The database suggests that the unknown compound that we have investigated may be a cinchona alkaloid—one of a series of natural products that are from the bark of the Cinchona tree.

In addition to proposed chemical structures, the database also provides the user with additional information about the compounds such as chemical name, molecular formula, exact mass, literature references, etc. The literature references can be particularly useful in the case where the user wants to research a proposed compound even further.

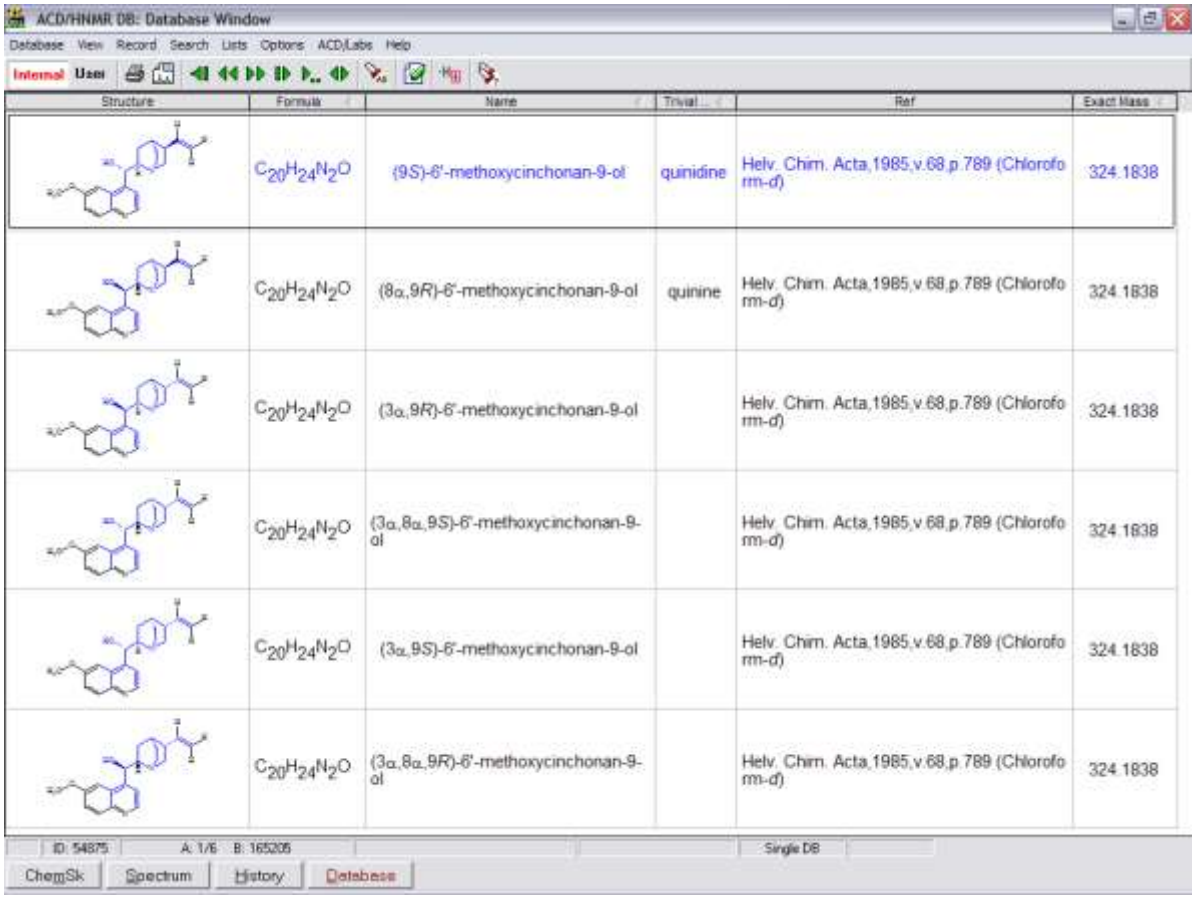

***Figure 5**: The results of the substructure search (shown in blue). Please note that 6 proposed chemical structures resulted when further searches by a molecular formula query filtered down the initial 15 hits.*

As mentioned in the previous section, these dereplication methods strive to accelerate and improve the process of structure identification as much as possible. In the end however, it is up to the user to employ chemical knowledge to determine the true identity of the unknown compound. In some cases this may require peer-review or more NMR data. As mentioned in the flow chart in Figure 1, in cases where this dereplication workflow fails to provide viable suggestions for previously identified structures, ACD/Structure Elucidator Suite can be used for a complete Computer-Assisted Structure Elucidation (CASE) from the experimental data available.

Conclusion

The isolation and characterization of natural products has historically been a tedious and time-consuming effort. Now, more than ever, significant strides in dereplication are being made to accelerate this process. The methods outlined above can assist a natural product chemist in quickly identifying substances that have already been studied. In less than 15 minutes on average, the software tools presented here can help identify potential structure matches from a minimal amount of spectral data. In doing so, this method can improve the efficiency of a group's dereplication efforts and accelerate the identification and development of potential drug candidates.

References

1. Williams, RB, O'Neil-Johnson, M, Williams, AJ, Wheeler, P, Pol R, Moser, A. Dereplication of natural products using minimal NMR data inputs. Organic & Biomolecular Chemistry, 2015, 13, 9957-9962. http://pubs.rsc.org/en/content/articlelanding/2015/ob/c5ob01713k
2. ACD/H/C/N/F/P NMR DB. http://www.acdlabs.com/products/adh/nmr/nmr_pred/. December 13, 2015.
3. ACD/Structure Elucidator. http://www.acdlabs.com/elucidator/. December 13, 2015.
4. Chapman & Hall/CRC Dictionary of Natural Products. http://www.chemnetbase.com/.
5. American Chemical Society: Scifinder. https://scifinder.cas.org/.
6. Royal Society of Chemistry: ChemSpider http://www.chemspider.com/, accessed May 2015.
7. ACD/Spectrus Processor. http://www.acdlabs.com/products/adh/spectrusprocessor/. December 13, 2015.
8. Moser A, Elyashberg ME, Williams AJ, Blinov KA, DiMartino JC. **Blind trials of Computer-Assisted Structure Elucidation software**. Journal of Cheminformatics, 2012, 4:5. http://www.jcheminf.com/content/4/1/5/abstract