Efficient Approaches for Addressing Spectral Ambiguities in Computer Assisted Structure Elucidation (CASE) Systems

Dimitris Argyropoulos, Rostislav Pol, Mikhail Elyashberg and Sergey Golotvin

Advanced Chemistry Development, Inc. (ACD/Labs), 8 King Street East, Toronto, ON, M5C 1B5, Canada

Introduction

Computer Assisted Structure Elucidation (CASE) expert systems (ES) have significantly facilitated the *de novo* structure elucidation of natural and synthesized compounds, especially in cases where the traditional (manual) methods fail. Current ES are based on NMR data, given that the molecular formula (MF) has been determined by HR-MS. Present CASE systems offer a few advantages:

1) They deliver all possible structures deduced from a given set of NMR data;

2) The fast empirical methods for NMR prediction allow selection of the most probable structure;

3) DFT based chemical shift calculations can confirm the selected structure;

4) They are now capable of suggesting a 3D model of the elucidated structure.

However, expert systems are still susceptible to a series of limitations which are mainly associated with the ambiguity in NMR spectra. Experimental ambiguity can result from low resolution 2D spectra, uncertainty in the hybridization state of carbon nuclei or atoms with possible variable valences (e.g., N and/or P) in molecules. Ambiguity can also arise from missing correlations in HMBC or COSY, often evident in hydrogen deficient molecules.

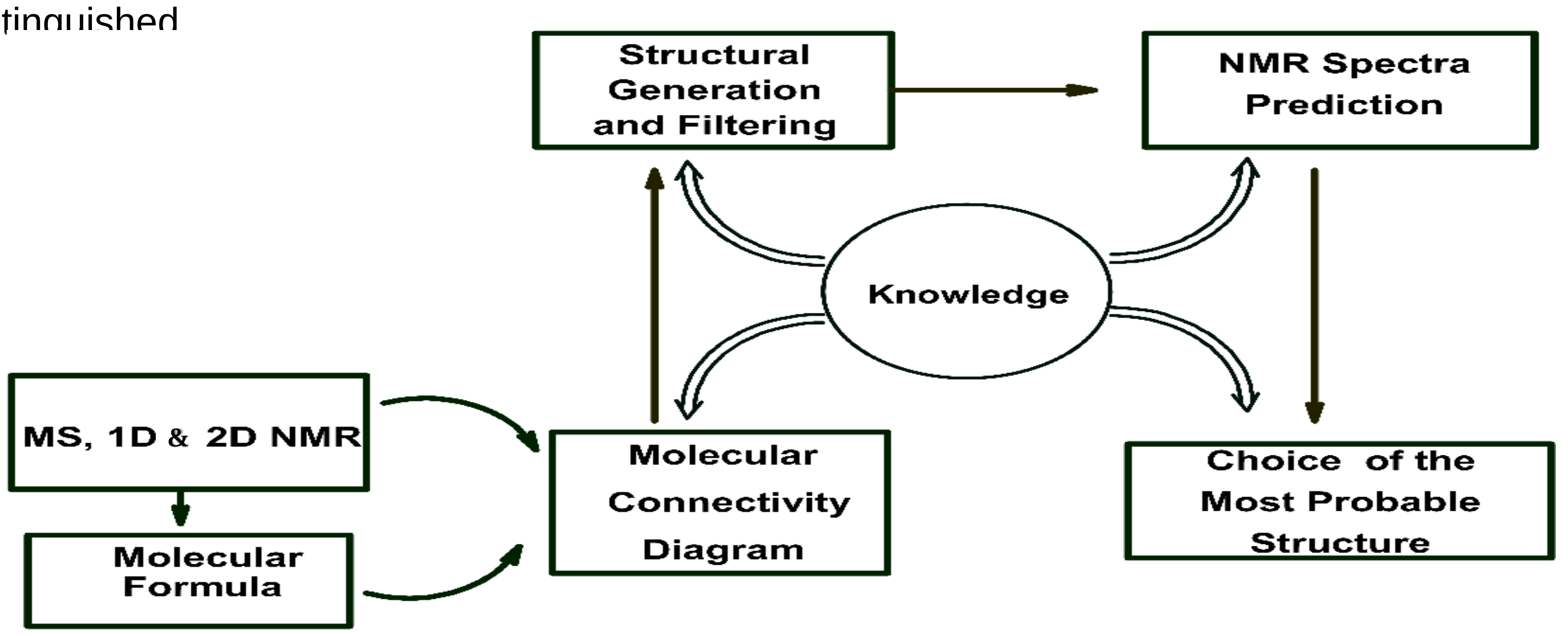
In order to remove the uncertainty, additional experiments are usually carried out however, such solutions are not always useful. In such cases, the only solution is the exhaustive investigation of all alternatives resulting from the presence of any ambiguity which could significantly increase the structure generation time (t_g).

Here we present approaches that enable structure elucidation when the data contains many ambiguous assumptions. Examples are presented, and the strengths and the limitations of each approach are discussed

Methods

In order to discuss the main principles on which CASE is based, we use the ACD/Structure Elucidator, as one of the most advanced examples in the market[1]. A simplified flow diagram of Structure Elucidation is presented in Figure 1.

Once the MF and the series of NMR spectra (1D ^{13}C and ^{1}H, HSQC and HMBC, ^{1}H-^{15}N COSY or almost any other 2D spectra) are fed into the program, the following steps in the structure elucidation can be distinguished


Structural Generation and Filtering

NMR Spectra Prediction

Knowledge

MS, 1D & 2D NMR

Molecular Formula

Molecular Connectivity Diagram

Choice of the Most Probable Structure

Figure 1. A simplified principal flow diagram of Structure Elucidator

Once the peaks are identified and peak-picked, the 2D correlations are translated into connectivities between the corresponding skeletal atoms. These connectivities are then used to automatically create a Molecular Connectivity Diagram (MCD), which displays the structural blocks and heteroatoms, and the possible attached hydrogens. The program annotates the structural blocks with ^{13}C and ^{1}H chemical shifts and atom properties (possible hybridization, forbidden or obligatory neighborhood with heteroatoms, etc.). The MCD information, which can also be edited by the operator, is then read by the program.

To select the most probable structure, NMR chemical shift prediction is performed and combined with spectral and structural filtering of the generated structures. Subsequently, the output file is ranked based on the average deviation calculated for each structure, and further filtering is performed using empirical methods based on the HOSE codes and Artificial Neural Networks prediction approaches[3].

Results

CASE Applications in Resolving Ambiguities

Example 1 | "Nonstandard" correlations in HMBC and COSY

Correlations, that exceed three bonds (i.e., $n>3$ for $^{n}J_{HH}$ and $^{n}J_{CH}$) are referred to as *nonstandard correlations*, NSCs. If the 2D NMR data contain undistinguishable *standard* and *nonstandard* correlations, the total set of "axioms" derived from the 2D NMR data will become logically *contradictory* and a correct structure cannot be deduced. In this case Fuzzy Structure Generation (FSG) is used to identify the presence of NSCs. Fuzzy Structure Generation will extend the observed correlations one bond at a time, until a structure is generated. This way one can resolve the contradiction automatically.

To illustrate the applicability of CASE to resolve this issue, Figure 2 shows a proposed structure for a product with 9 NSCs in HMBC (two *intense* $^{5}J_{CH}$ cross peaks), preventing structure elucidation using a traditional approach.

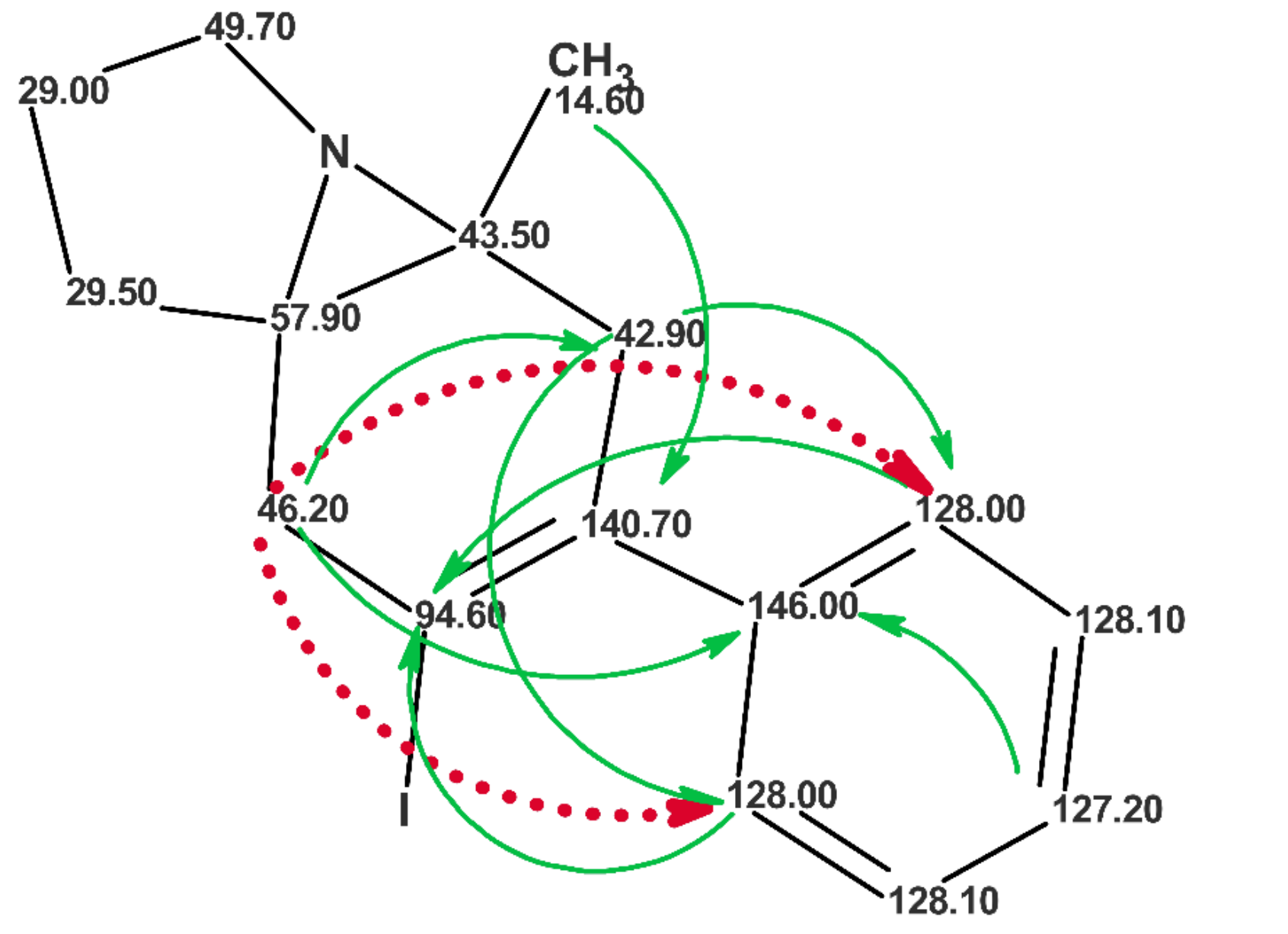


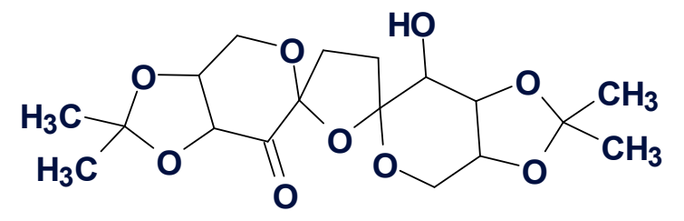
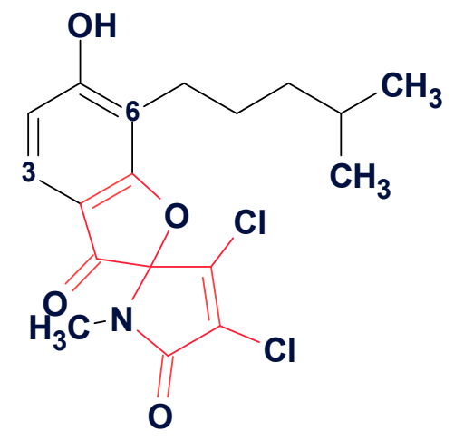
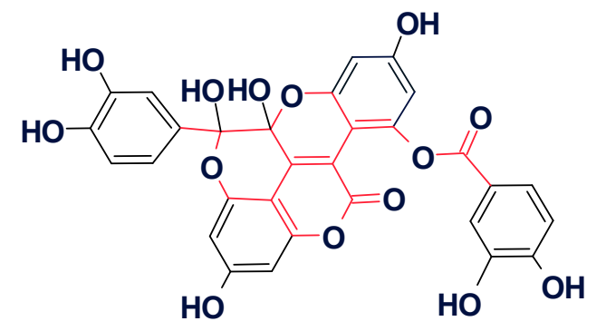
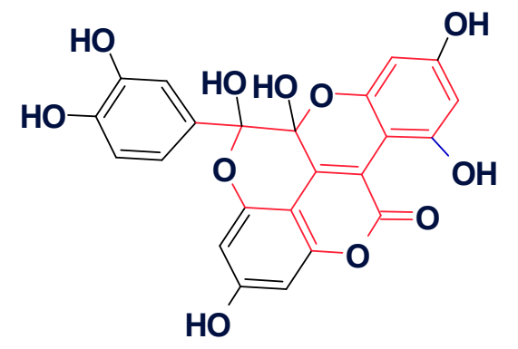
Figure 2. Suggested structure of a reaction product. The arrows show the ^{1}H -^{13}C HMBC NSCs. The connectivities from 46.2 to 128.0 marked by red dotted arrows correspond to $^{5}J_{CH}$.

The MF, 1D ^{13}C, HSQC, ^{1}H–^{13}C HMBC, and ^{1}H-^{15}N HMBC data were fed into Structure Elucidator , three structures were generated by Fuzzy Structure Generation in 13 min and the correct structure was ranked first. When 1,1-ADEQUATE correlations were also added to the 2D NMR data, only the correct structure was generated in 0.7s.

Example 2 | Hydrogen deficient molecules

In hydrogen deficient molecules (e.g., ratio of number of skeletal atoms to hydrogens >2), structure elucidation will become problematic. These molecules usually contain big "silent" fragments that are deprived from hydrogens, leading to an interruption in the correlation network.

Table 1. Examples of hydrogen deficient molecules of natural products whose structures were elucidated using CASE. The highlighted fragments in red are deprived of hydrogens.

Structure	Protocol
	$C_{18}H_{28}O_9$ R=1.3 k=64,372→4,755 t_g= 2m 30 s
	$C_{18}H_{19}Cl_2NO_4$ R=1.9 k=19,834→3 t_g= 15 m
	$C_{30}H_{18}O_{14}$ R =2.5 k = 176,400 → 42 t_g = 23 s
	$C_{23}H_{14}O_{11}$ R =2.5 k = 1,025,360 → 76 t_g = 6 min 40 s

INADEQUATE and 1,1-ADEQUATE spectra can help to overcome the deficit of structural information. If these information are not sufficient for structure generation within an appropriate time, use of molecular fragments greatly facilitates the solution. This is possible only if all ^{13}C experimental chemical shifts of the fragments are provided. The accommodation of one or more fragments within a set of connectivities derived from the 2D NMR data is also possible through the new algorithms. These fragments can frequently be found in the Structure Elucidator fragment library (ca. 2,400,000 entries). Structure Elucidator has been successfully used for structure elucidation of hydrogen-deficient natural products, few of which are presented in Table 1.

References

1. Steinbeck, C., Recent developments in automated structure elucidation of natural products. *Nat. Prod. Rep.* **2004,** 21, 512-518.
2. Blinov, K. A.; Smurnyy, E. D.; Curanova, T. S.; Elyashberg, M. E.; Williams, A. J., Development of a fast and accurate method of ^{13}C NMR chemical shift prediction. *Chemom. Intell. Lab. Syst.* **2009,** 97, 91-97.
3. Smurnyy, Y. D.; Blinov, K. A.; Churanova, T. S.; Elyashberg, M. E.; Williams, A. J., Toward More Reliable 13C and 1H Chemical Shift Prediction: A Systematic Comparison of Neural-Network and Least-Squares Regression Based Approaches. *J. Chem. Inf. Model.* **2008,** 48, (1), 128-134.

Example 3 | Ambiguous 2D NMR data

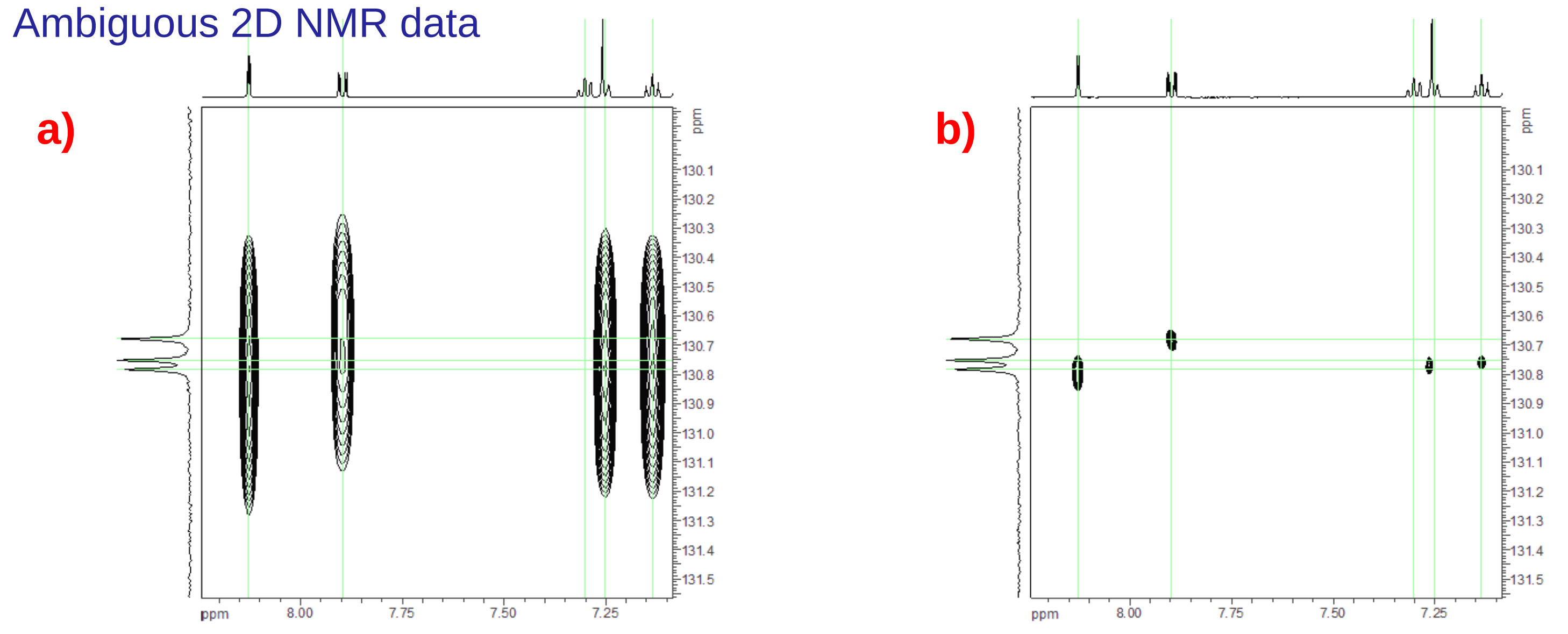


Figure 3. a) HMBC spectrum of a compound indicating the ambiguity arising from low resolution in F1. The ^{13}C signals at 130.7 ppm are not resolved. b) Band-selective HMBC spectrum of the same sample. The spectrum is recorded with the same number of increments and total acquisition time, but only between 110 and 150 ppm in F1. Peaks in this region can be unambiguously assigned to the proper carbon signals.

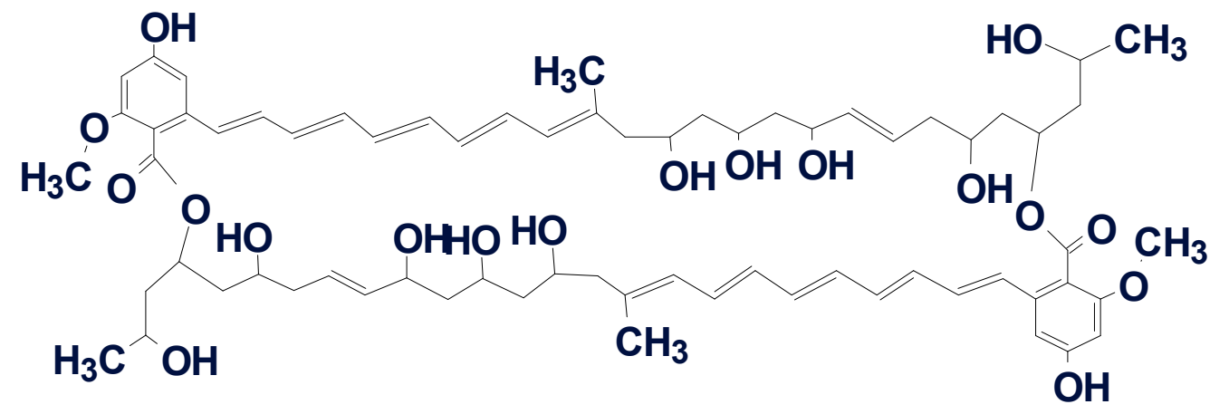
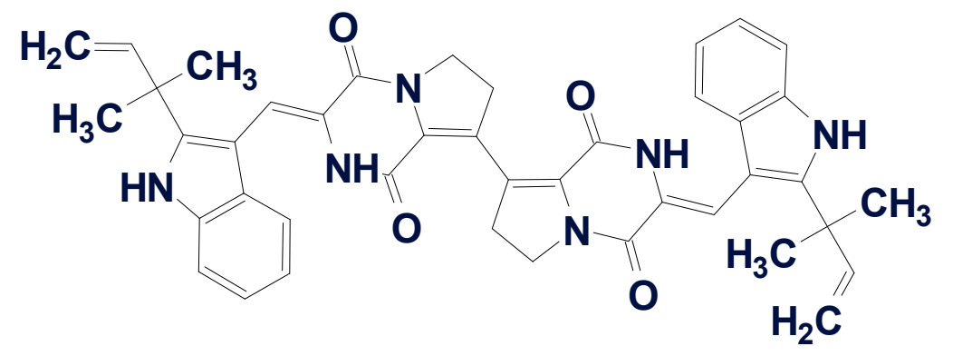
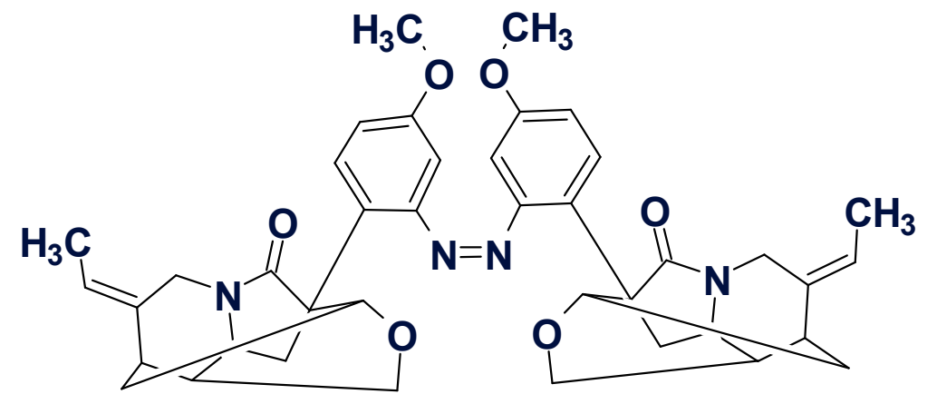
Low resolution 2D spectra cause spectral ambiguities for CASE. For example in Figure 3.a, it is impossible to confirm the correlations of the protons with either of the carbon signals, or even to both. Under these conditions t_g will be greatly increased to investigate all possible options.

This can be overcome by using modern NMR experimental techniques, such as band-selective versions of the traditional 2D experiments that record a small part of F1 with the same number of increments. Therefore, a much higher resolution potentially allows assignment. Alternatively the 2D experiment can be recorded using Non-Uniform Sampling (NUS) and processed to the required resolution.

Example 4 | Symmetric molecules

An enhanced self-adaptive algorithm has been developed that identifies features of molecular symmetry in 2D NMR data and adjusts itself to the generation of symmetric molecules. With this algorithm, the processing time for the generation of symmetric molecules from 2D NMR data is reduced to the typical order for unsymmetrical molecules. Some examples of symmetric molecules elucidated from 2D NMR data are shown in Table 2.

Table 2. Examples of symmetric molecules elucidated by Structure Elucidator

Structure	Protocol
	$C_{68}H_{92}O_{18}$, R =0.9 k =96→6 t_g =3 min
	$C_{42}H_{40}N_6O_4$, R =1.3 k =720→21 t_g =35 min
	$C_{40}H_{44}N_4O_6$, R =1.1 k =384→93 t_g =47 s

Multi-threaded Processing

Despite these advancements, there are still many cases where nothing can be done regarding the ambiguity. During the structure generation, several options need to be explored, resulting in the generation of several MCDs, one for each possible ambiguity. The problem is split into several smaller problems which individually do not have any ambiguities. These constituent problems were traditionally solved sequentially, but modern computer hardware allows for them to be solved in parallel, distributing them to different CPU cores. Depending on the time required to solve each problem this method can enhance the elucidation time, as shown in Table 3.

Table 3. Examples of multi-threaded processing applied

Compound	Molecular Formula	Speed gain
Retrorsine	$C_{18}H_{25}NO_6$	1.3
Artemisinine	$C_{15}H_{22}O_5$	4
Gymnopalyne A	$C_{12}H_7O_2Cl$	3.4
Schilancitrilactone A	$C_{29}H_{36}O_{10}$	1.8
Strynuxline A	$C_{23}H_{24}N_2O_5$	2
Epipancratitastin	$C_{14}H_15NO_8$	9
Eryngiolide A	$C_{20}H_{30}O_8$	9.5

Conclusions

Spectral ambiguity will always be present because of the physics used in recording NMR experiments. Although it may be a rather daunting hurdle to overcome in some cases, the use of newer NMR experiments and advanced structure generation algorithms, together with taking advantage of modern computer hardware can make its effect less pronounced.

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