Generating Unbiased Structural Alternatives for Automated Structure Verification

Sergey Golotvin, Rostislav Pol, Mikhail Elyashberg, Dimitris Argyropoulos and Karim Kassam
Advanced Chemistry Development, Inc., 8 King Street East, Suite 107, Toronto, Ontario, Canada

ACD/Labs

Introduction

Automated structure verification (ASV) using NMR data is gaining acceptance as a routine application for the qualitative evaluation of large compound libraries produced by synthetic chemistry. A proposed structure is confirmed if it fits a number of conditions from 1D ^{1}H NMR data [1] or from a combination of 1D-^{1}H and HSQC data.

Although it is easy to conclude that a proposed structure does not pass this "NMR filter" it is not trivial to guarantee the opposite—that a proposed structure passing this filter is the correct one. There is always a possibility that an isomeric structure fits the same NMR data better than the proposed structure. One way to decrease the possibility that false structures will pass, is to use as many NMR experiments as possible for structure validation (1D-^{13}C, HMBC, etc.). Another method [2] is to bring in and simultaneously verify several structures that are isomeric to the proposed. This combined concurrent structure verification (CCV) enables structural differences to be highlighted through the gradual tightening of NMR prediction tolerances until only a single structure remains. This CCV method certainly improves the confidence in the structure that passes while other similar structures don't, and provides a warning when one of the alternate structures also becomes consistent with NMR data. However, the CCV procedure does have a limitation – it only checks a few isomers that are selected either by a human expert or by an automatic algorithm (see Fig 1). In either case, the experiment presents a "bias" through the proposed structure itself, and as a result, other less related structures that could also fit the NMR data simply don't get checked.

One may argue that the probability of discovering a different structure as a result of synthesis with known starting materials is very low. Still, there is always a chance that an incorrect reagent was used or an undocumented impurity was present in a supplier product. One may want to take proper care when establishing the true identity of the sample.

Method

Here, we present a method for "unbiased" alternative structure generation for ASV based on the structure generator that is commonly used in a CASE system[3]. The method requires a proposed structure and an NMR dataset that is representative enough to support structure elucidation from it (typically ^{1}H, ^{13}C, HSQC and HMBC spectra). The NMR data are automatically processed and the proposed structure is automatically assigned (this is a necessary condition to reduce the possibility of false positives). The next step is the generation of the molecular connectivity diagram, followed by structure generation where all possible isomers fitting the NMR dataset are produced. Finally, the resultant structures are ranked based on prediction deviation and/or complex match factor. This scheme (Fig 1.) should enable the user to verify the correct structure starting from any arbitrarily chosen structure that is consistent with the NMR data.

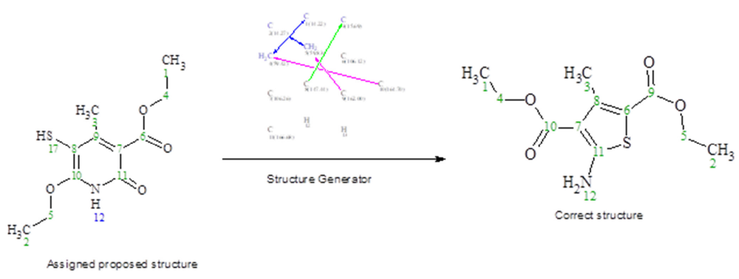


Fig 1. UBV scheme for structure generation

The main UBV advantage is that all possible structures compatible with NMR data are generated. A disadvantage, however, is that the generation process maybe lengthy. However using prior information like reliably present fragments (see below) can dramatically reduce the generation time.

What is the difference between UBV and structure elucidation? UBV has a lot of common with structure elucidation. But it is easier to use since the software performs all necessary checks for data consistency during the auto-assignment stage.

Experimental

Our primary goal was to study the ability of UBV to deal with false structures that were consistent with the NMR datasets but could not be generated and checked during the standard CCV procedure. To ensure the latter condition, one needs to take structures that could not be obtained from the proposed one by the following CCV rules:

1) move substituents along a ring
2) move substituents along a chain
3) Keep the number of C, CH, CH_2 and CH_3 the same as in the proposed structure

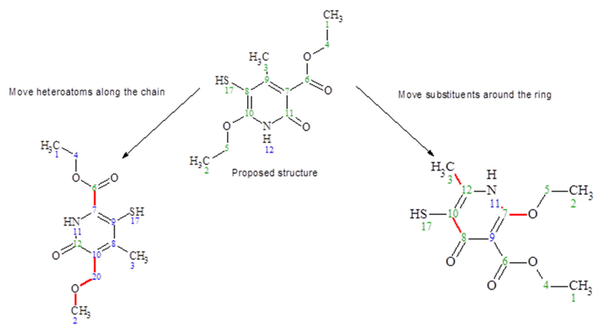
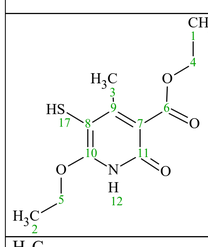
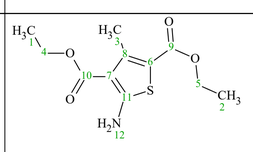
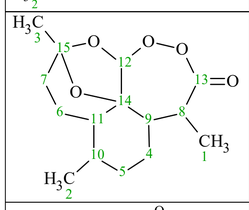
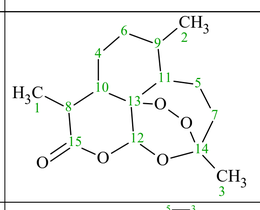
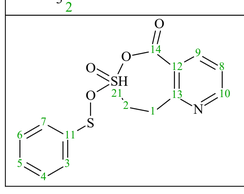
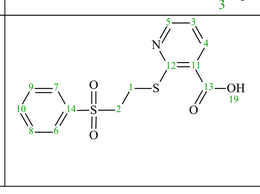


Fig 2. Automated (optional) isomer generation scheme of CCV. CCV advantages: isomer generation is very fast. Disadvantage: its ability to discover correct structure is limited by the choice of the proposed structure.

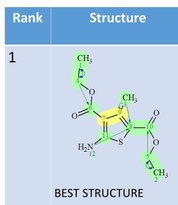
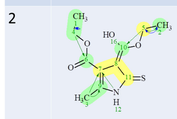
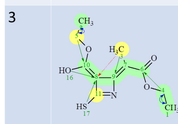
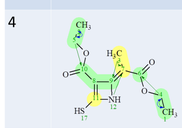
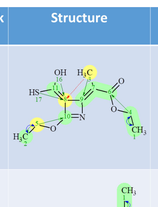
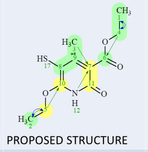
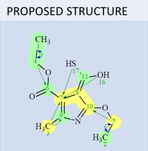
In practice, the following procedure was used for obtaining false structures that meet the above requirements. The structure elucidation procedure was run on NMR datasets of some 30 commercially available compounds and the resultant structures were subjected to ASV. For 3 of the samples, it was possible to find at least one false structure that passed ASV and was not related to the proposed structure as dictated by the rules described on Fig 2. These selected incorrect structures were then set as the proposed structures for the UBV procedure (see the Table 1)

Table 1 Proposed incorrect structures and correct structures as discovered by UBV.

Proposed incorrect structure	Generated "best structure"	Total number of considered structures
		339
		277
		218

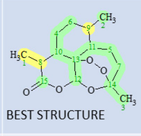
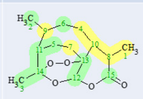
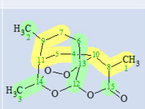
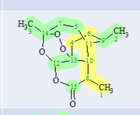
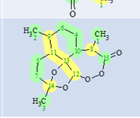
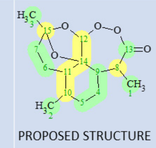
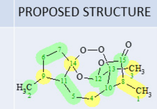
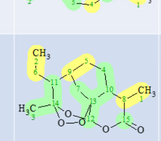
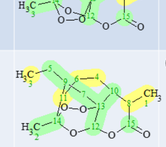
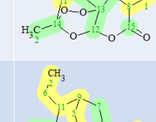
1st Test - Run the Proposed Structure through ASV with MF=0.83 using ^{13}C, ^{1}H, HSQC, HMBC and COSY

Then run UBV with no fragments defined, total time ~30mins, total number of unique structures generated =339

Rank	Structure	MF	dN(^{13}C+^{1}H), ppm*	MF(CV)
1	 BEST STRUCTURE	0.91	2.836	0.81
2		0.83	3.166	0.27
3		0.87	3.27	0.72
4		0.85	3.40	0.61
5		0.85	3.46	0.70
6	 PROPOSED STRUCTURE	0.83	3.46	0.54
7		0.89	3.60	0.75

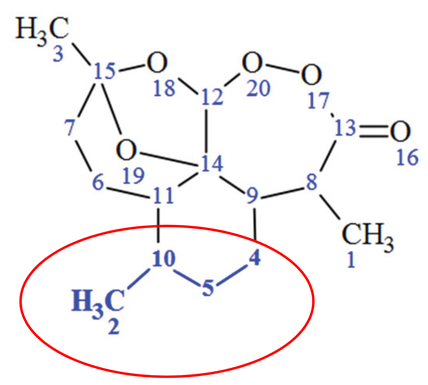
2nd Test - "Artemisinine" NMR data: ^{1}H,^{13}C, HSQC, COSY and HMBC – all processed automatically

- MF of the Proposed structure: 0.75
- Run CCV with 10 automatically generated isomers, the result is MF(PS)=0.75, all others are zero.
- Run UBV on the PS. Total time 5min 40sec, total number of unique structures generated = 277

Rank	Structure	MF	dN(^{13}C+^{1}H), ppm	MF(CV)
1	 BEST STRUCTURE	0.83	3.71	0.79
2		0.78	4.78	0.59
3		0.77	5.27	0.67
4		0.68	5.55	0.56
5		0.68	5.62	0.0
6	 PROPOSED STRUCTURE	0.75	5.69	0.64
7		0.75	5.79	0.64
8		0.68	5.80	0.0
9		0.71	5.84	0.0
10		0.69	5.94	0.26

After these top 10 generated structures have been selected for concurrent verification, only the Artemisinine structure passed this test.

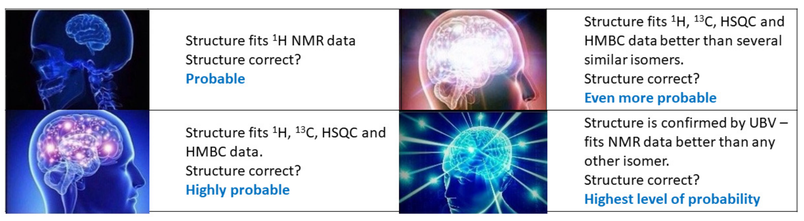
3rd Test - Run UBV on PS but define the fragment below as obligatory



The UBV generation time was markedly reduced from 4min 50sec to 15sec, the total number of structures generated was also reduced to 57. The best structure again was that of Artemisinine while the PS was ranked 4th. The defined fragment was incompatible with many of the structures generated before thus dramatically reducing the generation time.

Discussion/conclusions

Increasing the confidence that the proposed structure is correct:
"I have a structure that fits ^{1}H, ^{13}C, HSQC, HMBC .. spectra. Is this structure correct?"



A new method, Unbiased Verification (UBV), to generate and explore all possible structural alternatives compatible with NMR datasets is presented. The method can be applied to any proposed structure that passes standard NMR verification and can potentially be falsely considered as positive. The method was tested on several datasets with deliberately proposed incorrect structures where it automatically discovered the correct structures. Any potentially excessive calculation times can be reduced by introducing reliable structural information and further work is planned to achieve this without user intervention. We consider this Unbiased Verification method to be the ultimate tool for resolving potential false positives in Automated Structure Verification workflows.

References

1. Golotvin, Sergey S., Vodopianov, Eugene, Pol, Rostislav, Lefebvre, Brent A., Williams, Antony J., Rutkowske, Randy D., Spitzer, Timothy D., Mag. Res. Chem. 45(10), 803-813, 2007.
2. Golotvin, Sergey S., Pol, Rostislav, Sasaki, Ryan R., Nikitina, Asya, Keyes, Philip, Mag. Res. Chem. 50(6), 429-435, 2012.
3. Elyashberg, Mikhail E, Williams, Antony J., Blinov, Kiril.A., "Contemporary computer-assisted approaches to molecular structure elucidation", RSC, Cambridge, 2012

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

Visit ACD/Labs at our booth