Automated Structure Verification: What are the Right Experiments and Processing?

Sergey Golotvin[1], Patrick Wheeler[1], Phil Keyes[2], Rostislav Pol[1] and Gerd Rheinwald[1]
[1]Advanced Chemistry Development, Inc. (ACD/Labs), 8 King Street East, Suite 107, Toronto, ON, Canada, M5C 1B5
[2]Lexicon Pharmaceuticals, Princeton, New Jersey, USA

Introduction

Standard chemical structure characterization regularly employs a variety of 2D NMR techniques. However, past practice for the computer automation of this technique, Automated Structure Verification (ASV), primarily employs either 1D ^{1}H NMR only, or a combination of 1D ^{1}H NMR and 2D ^{1}H-^{13}C HSQC. [1] Recent development makes the inclusion of a wide array of experimental data possible in fully automated structure verification work. The inclusion of expanded data types supports more accurate structure verification, decreasing the likelihood that false structures may pass through a verification process.

Recent experimental work has provided a rich array of experimental data on a large variety of structures for chemical samples that are derived from several sources. Included are 1D ^{1}H, 1D ^{13}C, 1D ^{13}C DEPT, ^{1}H-^{13}C DEPT-edited HSQC, unedited ^{1}H-^{13}C HSQC, COSY, TOCSY, and HMBC data. Coupling the analysis of such data with the ability to create spectroscopically relevant challenge structures [2] enhances the certainty of the chemist that they have synthesized the correct structure, and the confidence with which an organization can assume that the structure of any component in its library is completely correct.

Analysis of this variety of data sets helps to establish the most efficient experimental processes to ensure that correct structures are rapidly recognized, while incorrect chemically relevant structures are flagged for failure or for further analysis.

Here we present an analysis of several different correlation techniques in order to better understand the value of various NMR experiments in ASV work.

Experimental

A wide variety of data sets were collected on an array of 51 compounds at 700 MHz on a Bruker Avance III spectrometer using a 5mm TCI-cryoprobe at 300K. The experiment types included, 1D ^{1}H and ^{13}C, DEPT135, 2D COSY, TOCSY, HSQC, HSQC-DEPT, and HMBC. Some ^{13}C DEPT and ^{13}C spectra were collected on 400 MHz Bruker Avance II and Avance I spectrometers running TopSpin 2.1 and 1.3 respectively with a 5mm BB probe at 300K. All 1D ^{1}H and 2D data acquisition and most 1D ^{13}C and DEPT data acquisition were run on samples prepared at 10uM concentration generally in DMSO-d_6 using 3mm and 5mm tubes. A small number of samples for 1D ^{13}C were prepared at concentrations between 20 to 35uM. All initial processing was performed via ICONNMR. HSQC-DEPT spectra were manually phased according to the negative phasing of methylene convention when found inverted in automated processing.

Automated verification was carried out using ACD/Automation Server V.2015 with a Match Factor Cutoff of 0.69, chemical shift tolerances of 1.3 and 13 ppm for ^{1}H and ^{13}C respectively. For the sake of time, neural net predictions were used for all verifications. For multi-bond experiments all peaks were accepted that are by integral more than 0.67 of the median integral in each spectrum.

Multiple challenge structures (up to 10) were generated that demonstrate spectroscopic similarity. The algorithm allows movement of hetero-atoms and chain attachments up and down chains and around rings, but not changing molecular formula, numbers of CH, CH_2, or CH_3 groups, or ring dimensions according to rules previously suggested. [2] A Multi-Step Verification approach was used; in this case, when more than one structure passes a verification test, the chemical shift tolerances are tightened by 10% per iteration over a course of up to 5 iterations to determine whether any structure can be differentiated.

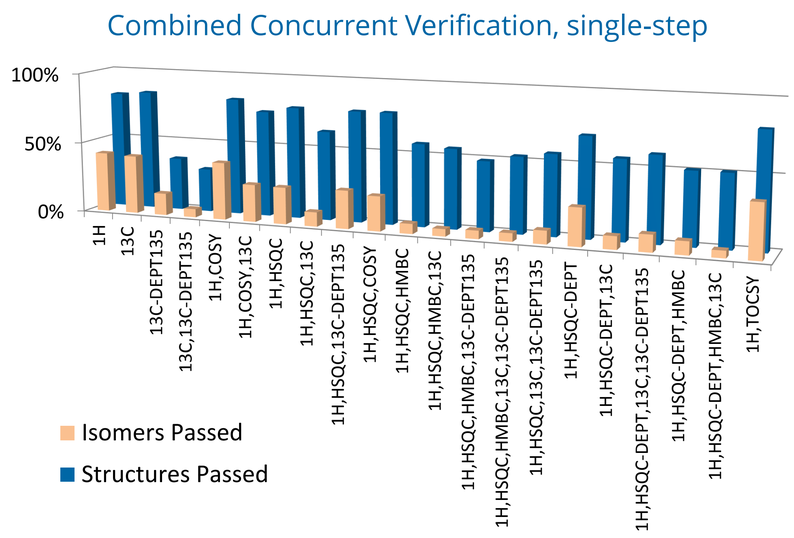

Figure 1: A table of the Combined Concurrent verification results from one compound set. Note the large number of passed isomers when no Multi-Step Verification is employed.

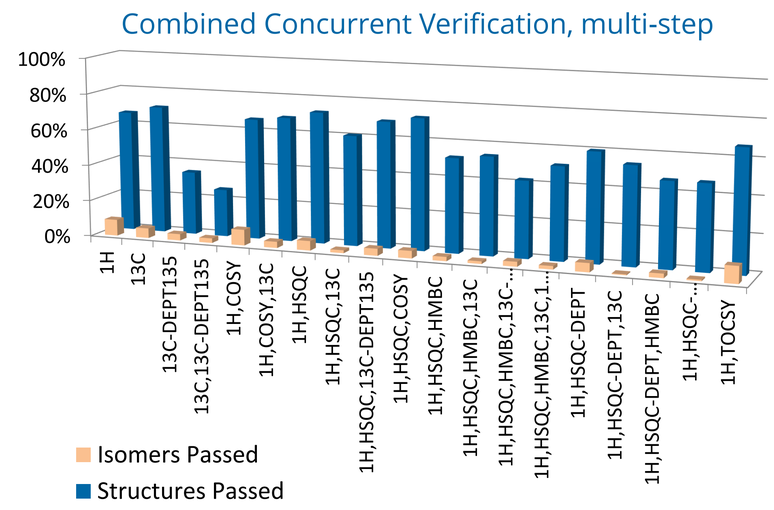

Figure 2: A table of the Combined Concurrent verification results from one compound set, employing Multi-Step Verification. Note the very substantial change in differentiation between Correct Structures and proposed alternative isomers, but the still-high false positive rate for ^{1}H only spectra.

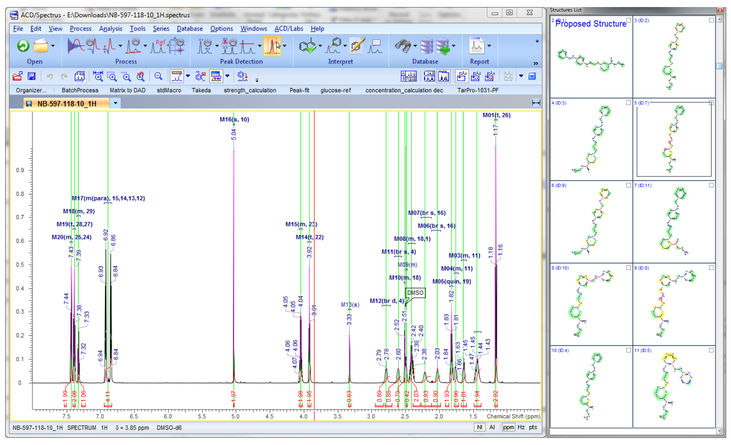

Figure 3: An example set of data including challenge structures created by the software. In this case, using only 1D ^{1}H data, the Proposed Structure and five alternatives all met the passing criteria.

Experiment	Experiment Time (minutes)	RDP
^{1}H	1.3	1
^{13}C	50	1.72
^{13}C-DEPT135	14	1.31
^{13}C, ^{13}C-DEPT135	64	1.38
^{1}H,COSY	10.5	1.03
^{1}H,COSY, ^{13}C	60.5	2.68
^{1}H,HSQC	15.3	1.85
^{1}H,HSQC, ^{13}C	65.3	**5.09**
^{1}H,HSQC, ^{13}C-DEPT135	29.3	2.45
^{1}H,HSQC,COSY	24.3	2.36
^{1}H,HSQC,HMBC	140.3	2.99
^{1}H,HSQC,HMBC, ^{13}C	190.3	**5.17**
^{1}H,HSQC,HMBC, ^{13}C-DEPT135	154.3	2.18
^{1}H,HSQC,HMBC, ^{13}C, ^{13}C-DEPT135	204.3	3.78
^{1}H,HSQC-DEPT	12.3	1.64
^{1}H,HSQC-DEPT, ^{13}C	62.3	**15.52**
^{1}H,HSQC-DEPT,HMBC	137.3	2.64
^{1}H,HSQC-DEPT,HMBC, ^{13}C	187.3	**8.81**
^{1}H,TOCSY	19.3	0.95

Table 1: List of experiments used, with their expected relative spectrometer time in this set of high s/n experiments, and their Relative Discrimination Power.

Results and Discussion

The NMR data acquired for the work with the structures was of very high quality, so that factors of s/n or resolution could be eliminated from consideration. Rather, this becomes a strict test of automatic assignment algorithms. Multiple combinations of data were considered for use in verification in an effort to determine the most appropriate approach. Not all data types were available for all structures, so only applicable combinations were considered for each structure. In general, the parameters were selected according to a false positive tolerant strategy. In this case, the parameters favor the passing of structures even where there may be some inconsistencies. This mimics a real-life situation in which there is a strong presumption that submitted structures will be correct, and only serious problems will be flagged.

However, despite the use of advanced high-quality ^{1}H and ^{13}C chemical shift prediction data, in conjunction with combined and multistep verification results, the difficulty for verification systems to differentiate similar isomers is apparent. The data shows that even when leveraging the computation power of advanced verification systems using prediction and assignment algorithms, the potential universe of constitutional isomers can easily provide alternative challenge structures that are also consistent with the data. In order to provide true confidence that a proposed structure is indeed highly likely to be the only authentic structure that the set of data supports, one needs to reduce the number of alternative isomers passing to a suitable figure, probably below 5%. Multi-step verification provides great assistance in this, but there is still demonstrated peril in the optimistic submission of final compounds supported solely on chemist verification of ^{1}H NMR and the presence of the parent ion in LCMS data.

The decreasing true pass rates when using long range correlation experiments such as the HMBC suggest that the scoring mechanisms require further development.

By considering the pass rate for a given experiment, or combined set of experiments, relative to a series of generated isomers used as challenge structures against a verification outcome, we can quantitatively establish a Relative Discrimination Power (RDP) for a data set. The reference benchmark used is a standard ^{1}H spectrum. By calculating a structural discrimination ratio where the number of passed results of a given set of correct structure/data pairs (true positives) is divided by the total number of isomers (negative control challenge structures per set of experimental data) that also passed (false positives) and subsequently using the resulting calculation of the ^{1}H data set against the structures as the benchmark, an RDP value can be determined to illustrate how much more effective a given set of experimental data can be at eliminating false negatives relative to other sets.

$$RDP_{exp} = [(TP_{exp}/FP_{exp}) / (TP_{H}/FP_{H})] / Time_{exp}$$

With this information, selection of the best combination of data can be considered based on efficiency factors that take into account total spectrometer time versus the resulting distinguishability of structures with a combined data set.

Conclusions

What is the best combination of experiments and computation methods for deriving proper True Pass rates, while also eliminating as many False Positives as reasonable?. First, we have created a new metric, the RDP, which assists in understanding quantitatively the value of different experiments in the verification process. Second, the use of Multi-Step Verification adds immense power to discriminate between proper structure and similar but incorrect isomers. Third, the greatest power is achieved only with the inclusion of high-resolution ^{13}C data, which is increasingly available. Lastly, there are not yet ideal algorithms for accommodating long-range correlations in verification.

Advanced Chemistry Development, Inc.
Tel: (416) 368-3435
Fax: (416) 368-5596
Toll Free: 1-800-304-3988
Email: info@acdlabs.com
www.acdlabs.com
Reprints:
conferences@acdlabs.com

References

1. Automated compound verification using 2D-NMR HSQC data in an open-access environment, Keyes, P., Hernandez, G., Cianchetta, G., Robinson, J., Lefebvre, B. Magnetic Resonance in Chemistry, Volume 47, Issue 1, pages 38-52, 2009.
2. Concurrent combined verification: reducing false positives in automated NMR structure verification through the evaluation of multiple challenge control structures, Golotvin, S., Pol, R., Sasaki, R., Nikitina A., Keyes, P. Magnetic Resonance in Chemistry, Volume 50, Issue 6, pages 429–435, 2012.