Automatic Structure Verification (ASV) as an AI wizard: the milestones met and the challenges looming ahead

MESTRELAB RESEARCH
NMR Solutions

Carlos Cobas[1], Felipe Seoane[1], Stanislav Sýkora[2]

[1] Mestrelab Research, Santiago de Compostela, Spain; carlos@mestrec.com
[2] Extra Byte, Castano Primo (Mi), Italy; sykora@ebyte.it

In collaboration with
www.ebyte.it

The community of NMR spectroscopists certainly does not need an explanation of the meaning of terms like “structure verification” or “structure elucidation” or, more generally, “interpretation of a set of NMR spectra”. Likewise, the addition of terms like “computer aided” or “automatic” of the previous terms is quite straightforward and intuitive. But is it really? Actually, from the algorithmic point of view, each of these terms means something considerably different (though related). For example, it may be not quite clear to those who are not actually involved in software development how radically different is a “computer aided” evaluation of a set of experimental, real-world data, and a completely “automatic” handling of the same set. Or how much does ASV differ (using the same set of data) from the task of discriminating between several molecular structures. Likewise, while everybody understands that software products addressing these tasks should help the spectroscopist to carry out his tasks, very few think that it will be ever possible to eliminate a major portion of his today’s everyday work. We believe that we have reached a point in the development of ASV where we do understand the real limits and can address the above distinctions and judge how far one can follow this road.

Here we address the above topics, and also share some of the achievements, milestones, victories, tricks, and – why not – errors, we went through during this intense four-year work. We believe that the data-evaluation wizard we taught to stand on its feet and walk, and which we are now teaching to run, will become much more than an NMR tool. Amazingly many questions had to be addressed to make it work: data reduction into easily manageable forms such as the global peaks list (GSD), combined with the elimination of many artifacts, detailed editing of the reduced data, massive “filtering” of useful information from the accidental, handling of real-world fuzziness and aspects of the data which still remain ill-defined, even after all the filtering, the way of incorporating specific know-how (in this case, “The NMR Book”), the way to ponder and combine fragments of information arriving from quite different quarters and based on different considerations, and a massive application of new algorithmic techniques suitable to mimic parallel “thinking” and even “intuition”.

Much of this experience, we believe, could be put to work also in areas other than NMR but that, of course, goes beyond this presentation. We do believe, however, that through the development of this kind of a wizard, NMR can significantly contribute to the understanding of how a real-world artificial intelligence should be built.

As an example what the wizard is doing during an ASV run, consider the standard 1H and the Edited-HSQC spectrum of Ramipril, one of the most common farmaceutical substances. The raw spectra, as well as the structural formula of the compound are shown on the right (Figures Row 1, left and right, respectively). The sample was a room-temperature solution in d_6-DMSO (di-methyl-sulfoxide).

It is evident that no software could ever proceed and evaluate such spectra as arrays of numbers with all their artifacts (such as receiver noise and baseline roll). Such a procedure would be extremely unwieldy, inefficient and inflexible.

The first step must be, unavoidably, a **reduction of the raw data** into lists of basic pertinent features which, in this case, are the **spectral peaks**. It is in fact a wise thing to make the code mimic, as much as possible, the time-honored methodology developed in describing and analysing the spectra «manually».

This is achieved using the procedure called GSD[1] (Global Spectrum Deconvolution) which:

-Discerns the individual peaks,
-Fast-fits each peak’s parameters (position, height, width, and position) trying to match the spectrum.
-Lists the results in a digital «table»

In this particular case the algorithm identifies 315 peaks in the 1D spectrum and 146 peaks in the 2D spectrum (see portions of the peak tables in Figures Row 2, left and right, respectively).

The main point in the present context, however, is not the data reduction to the «tabulated» form as such, but the fact that the procedure functions as a first «filter», separating potentially valuable information from undesirable and irrelevant one (noise, baseline). Naturally, this involves soft thresholding, for example to decide whether a noise excursion is a peak or not, or whether a peak shoulder qualifies as an extra peak. This introduces the first «level» of uncertainty – something an automatic wizard needs to wade through all the time. As it proceeds through the various levels, from bottom up, the uncertainty generated at lower levels must progressively decrease in importance and impact, while new uncertain aspects emerge. The fact that thresholding (a non-linear operation) is heavily involved is reminiscent of neural networks which teach us that massive combinations of nonlinear transfer functions is a pre-requisite to extracting meaningful information from seemingly random data.

The second step consists in automatic **editing** of the peak lists. The primary purpose of the editing is to analyse each peak, both separately and in relation to its neighbors, and decide whether it is a compound non-labile or labile peak, a reference peak, a primary or secondary solvent peak, or a 13C satellite (in 1H spectra), or an impurity or artifact peak. The results of the basic classification are displayed and color-coded for easy User inspection (Figures Row 3, left and right, respectively, for 1H and 2D peaks).

The third step, consisting in **application-dependent data-analysis** is actually inter-twined with the editing. It goes much further than what appears in the displays. For example, the wizard tries and groups the peaks into multiplets (in 1D spectra) or clusters (in 2D spectra), tries and guesses the total number of protons in the molecule and the way they are grouped together, builds internal data structures such as JC (J-correlation) lists of edited splittings, and uses them for various purposes such as multiplets purging and slicing. All this is done using what ever information is available. For example, the availability of a hypothetical molecular structure (guaranteed in ASV) is obviously of great help but, the wizard does not stop when the molecule is unknown, thus opening the door towards Automatic Structure Elucidation (ASE).

Again, all this work relies heavily on non-linear statistical approaches, such as smooth thresholding, democratic scoring, and veto scoring. For example, to decide whether a peak might belong to the primary solvent, we apply a scoring system with 13 independent «tests». Similar scoring systems are used to score on the congruence of each peak with the secondary solvent, on its being a labile, on its membership in a multiplet, etc etc. Multiplets and clusters, of course, have their own multiple scoring systems themselves. A typical ASV run can involve around 10000 individual scoring tasks, all of them are to some extent «fuzzy» and include non-linear operations.

When an HSQC spectrum is available, it is used in several steps, starting from the editing (to check on labiles and water) and proceeding with elementary and global 1H assignments, up to its own self-standing ASV scoring. This is in line with another principle we have learned: generate and use any information as soon as possible (the greedy nature that any AI wizard should possess).

The fourth step is the generation of possible **elementary assignments** (which assigns a nucleus to a 1H multiplet or a 2D cluster) and enumeration of **global assignments** (the set of all elementary assignments, one for each nucleus in the molecule). Needles to say, this involves more scoring and some extra nuances, but the details really exceed the possible scope of this presentation. The one thing to point out is that as the wizard proceeds along the steps (levels) almost all the scoring criteria become less generic and more specific to the application at hand (in this case incorporating more and more NMR-specific know-how). Another feature which was not yet pointed out are frequent loop-backs: an intelligent wizard must understand when it is in a blind alley and be able to get out of it!

The final step is, of course, **the actual ASV**. After all the «preparatory» analysis of all available data, the ASV is little more than just «putting it all together». For completness sake, we show what the wizard comes up with when applied to our example and using various strating sets of data.

When only the 1H spectrum is used, the results are those in Figures Row 4, left. The elementary assignments are visualized by means of numbered disks or circles. The color coding of the disks/circles corresponds to the degree of acceptability of the elementary assignment. Solid disks indicate «stable» elementary assignments, i.e., those which remain the same in all top enumerated global assignments, while circles indicate elementary assignments which do not have this property. The Verification results box shows the outcomes of several independent ASV tests, and their combined final outcome. Despite one red elementary assignment, the overall outcome is positive: the molecule might be compatible with the spectrum. In the same row, on the left, an expansion of the spectrum illustrate some of the complexities of the analysis (no room to go into details here).

When only the Edited-HSQC data are used (the last-but-one Figures Row) the final verdict is the same, and most (though not all) of the assignments are compatible with the 1H case.

When both sets of data are used (the last FiguresRow) the outcome is actually less affirmative that in any of the two obove cases. This is due to the fact that while for most nuclei the two spectra strengthen each other, there appears to be a «conflict» for one of the nuclei (the detail on the right shows that the cause might be an impurity multiplet). The more information is used, in fact, the more likely it is that the wizard will find out a real-life imperfection. For that reason, it must be taught a correct degree of «tolerance» in order to be really useful.

References:
[1] C.Cobas, S.Sykora, The Bumpy Road Towards Automatic GSD, Poster at 50th ENC, DOI 10.3247/SL3Nmr09.003.

Presented at 54th ENC, Asilomar April 14-19, Asilomar, Pacific Grove, CA, USA. DOI of this document: 10.3247/SL4Nmr13.002