Progress in Multi-Spectra Automatic Structure Verification (MS - ASV)

Stanislav Sýkora[1] and Carlos Cobas[2]

[1]Extra Byte, Via Raffaello Sanzio 22/C, Castano Primo (Mi), Italy I-20022; sykora@ebyte.it
[2] Mestrelab Research, Xosé Pasín 6, Santiago de Compostela, 15706 Spain; carlos@mestrec.com

DOI: 10.3247SL6Nmr17.003

Introduction

Automatic Structure Verification (ASV) is fast becoming an important part of NMR data evaluation software packages such as Mnova and others. Its basic goal is to answer, in a qualitative as well as quantitative way, the question

"*Is this molecular structure compatible with these NMR data?*"

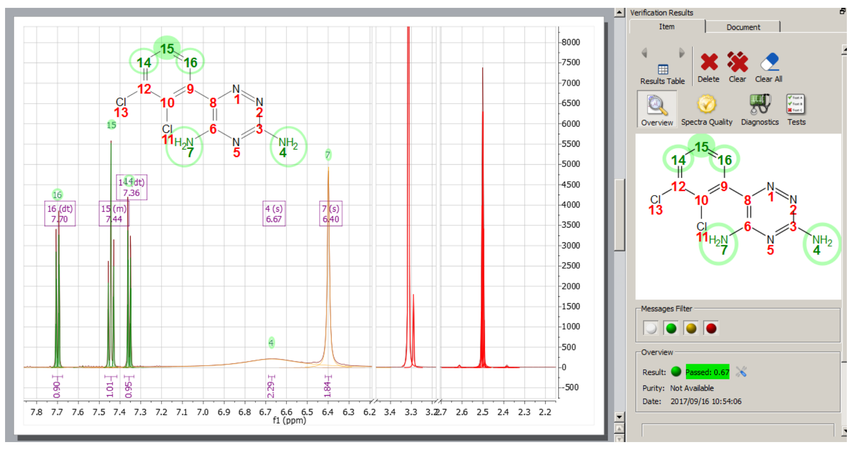
Naturally, the question one makes, the answer one gets. These things are fuzzy in logic because of imperfections and impurities in the spectra, nmr parameters prediction errors, solvent effects, etc. The complexity is close to that of an artificial intelligence, and the scoring is critically dependent on the query itself. But the problems extends beyond that, involving what one intends by "NMR data". As we know well, it is one thing to have a single 1H spectrum, for example, and another one to have a pair of spectra of different kinds, such as 1H and 13C, or 1H and HSQC, and a still more different one if the spectra (of presumably the same compound) include an arbitrary subset of any of the "200 and more" NMR experiments such as, for example 1H, 13C, HSQC, COSY, TOCSY, HMBC, ROESY, and others. How does one proceed with the automatic analysis in such cases? We all know that the likelihood of multiple "solutions" decreases sharply when one combines several spectra of different kinds. We also know that even 'human' analysis is in these cases anything but linear and standard: it requires multiple "passes" through the spectra and a non-trivial search for correlations (or lack of correlations) of various orders. We also expect that critical points (penalties) tend to accumulate when new data are added until they overflow a threshold (false negatives). Since no spectrum is perfect, how many spectra it takes before any structure gets ruled out? It implies that to reach a correct conclusion, there may be an optimal number of spectra; adding still more does increase the knowledge only marginally, while uncertainties continue to increase at least linearly.

Over the last few years we have built a considerable body of experience [1] with these problems, though we do not (and can not) claim that we have solved them yet. How to design a software machinery to iteratively analyze a set of any number of NMR spectra of various kinds (but presumably of the same compound, or better of the same sample)? It should extract and refine all information the spectra might contain, in a synergetic way, discard in an intelligent way all missing features and/or features in excess, and then knowledgeably score that information database against an assumed, hypothetical molecular structure. That, indeed, is a huge task which (in the opinion of one of the authors ☺) will keep to be tackled for the rest of this Century.

Defining the task, and the problems that unavoidably arise

Given a presumed molecular structure and a number of NMR spectra of a sample, one would like to know to which extent is the hypothesis (= the structure) compatible with ALL the available data, including chemical shift and coupling constants predictions, and what are the assignments of the expected equivalent nuclei to the individual experimental 1D *multiplets* and 2D *clusters*. These are actually two very different tasks, even conflicting tasks. For example, having multiple acceptable assignments reinforces the hypothesis that the molecule is 'correct' while it dramatically decreases the chance that any one assignment selected among the acceptable ones is fully correct. Yet, spectroscopists and chemists in general strongly want the two tasks to be carried out concurrently, and the results to make sense even for molecules which are 'incorrect', with maybe just fragments of them matching some of the spectral features.

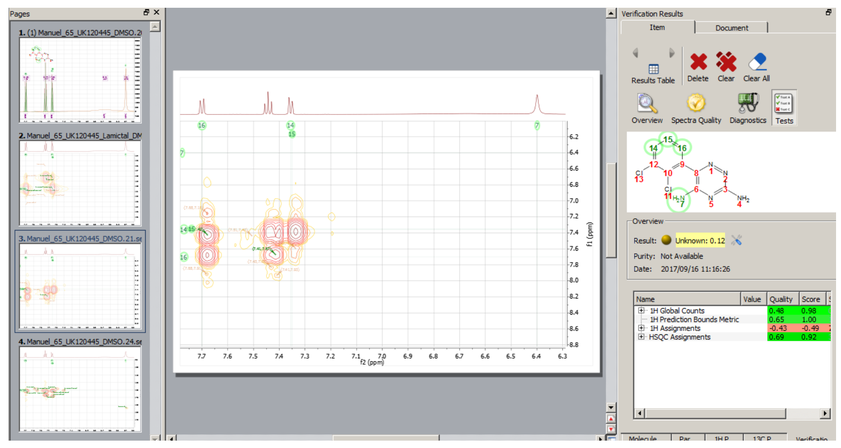
When dealing with 1D spectra the task can be reduced to looking for the best, non conflicting, assignment of each expected group of nuclei to one of the multiplets of the spectrum, respecting the multiplicities and integrals, and allowing for the overlap of the experimental multiplets. Not an easy task, to, but manageable. The result looks more or less like this:



However, when other spectra are added, particularly 2D, this approach is no longer convenient. Assignments need to be reinterpreted in terms of the numeric parameter values (shifts and coupling constants) in an expected spin system. This is the only possible approach that will guarantee coherence (synchronization) over ALL the spectra, but it is also a conceptually different from what one instinctively does in 1D-only setups. The 'solution' is then defined as the set of shift values for all distinct groups of nuclei in the molecule with all its symmetries. The values should be narrowed down as much as possible to match, simultaneously, all significant cross peaks in all 2D spectra, as well as all assigned multiplets in all 1D spectra.

This is a very complex combinatorial problem which, particularly in large molecules, may have an incredible number solutions (or, otherwise, not a single one) and there is no known algorithm that might help us with this – not with all the specific constraints NMR theory imposes on the various kinds of spectra.

In fact, when several spectra are included, the result, rather than reinforcing the structure confirmation, may look somewhat like this (1H, HSQC, COSY, HMBC):



So, how can a good molecule start looking WORSE when more spectra are added ???

In this case the molecule is small and simple, so the problem is not in the complex combinatorial nature of the task, but rather in the presence of many imperfections in the spectra. Imperfections such as:

- Missing peaks (e.g., in the central panel in the Figure on the right, one notices a missing pair of COSY cross peaks which, according to the cannons of NMR textbooks should be present.
- Significantly misaligned spectra (also present in this example).
- Impurities and other extra peaks which should not exist.
- NMR sequence artifacts (such as high order correlations) which may, but need not be present).
- Multiplets and clusters overlap.
- Etc. There is really a long list of such items and there is no room in a single poster to illustrate all of them.

The bottom line is that all these uncertainties need to be accounted for, in great detail, in any software attempting to tackle MS–ASV and, even so, it further complicates the combinatorial problem of finding a good 'solution'.

CONCLUSIONS

There is, and probably will never be, a perfect, mathematically proved solution to the problem of MS–ASV. Even without spectral imperfections, sample impurities, and instrumental and NMR sequence artifacts, it is a combinatorial problem of non-polynomial complexity (NP). Including the 'real-life disturbances' just listed, it becomes something like NP^{NP} problem.

Just like other well known NP problems (such as that of the travelling salesman, or the one of dividing N unequal tasks among M workers/processors), it has no rigorous mathematical solution provable to be the best one. On the other hand, NP problems are well known to admit many very different algorithmic approaches that generate reasonable practical solutions.

We are therefore most probably heading towards a long period of proliferation of multiple software MS–ASV approaches, both within each Software House (we went through several during the last year and our quest continues) and from different software developers (no doubt each will claim that theirs is the top killer). It is important to understand that this situation is a consequence of the objective mathematical nature of the task.

As an aside, we find interesting the finding that when more and more spectra are submitted for the same molecule, the results (both PASSes and FAILs) first become more affirmative but then start dropping sharply, unless the quality of the spectra is very high. This, to think about it, is quite logical: there is no perfect NMR spectrum, in particular not a 2D one. For any algorithm, the imperfections are confusing and generate scoring penalties. When the added spectrum does not contain any new information (one that can not be deduced from the other spectra), its contribution consists primarily just in adding a new load of penalties (when it comes to details, most finicky features of NMR lead to either a possible sharp penalty, or a very mild bonus; there is no way how a newly added spectrum might sharply boost the acceptance of a rejected molecule).

References

1. Cobas Gómez C.J., Bernstein M.A., Sykora S., An Integrated Approach to Structure Verification Using Automated Procedures, Chapter 12 in Structure Elucidation in Organic Chemistry: The Search for the Right Tools, edited by Bravo J., Cid M., Wiley-VCH 2015. DOI: 10.1002/9783527664610.ch12.

UNIÓN EUROPEA
Fondo Europeo de Desenvolvemento Rexional

XUNTA DE GALICIA
CONSELLERÍA DE INNOVACIÓN, INDUSTRIA E COMERCIO
Dirección Xeral de Investigación e Desenvolvemento

Developed in collaboration with www.ebyte.it