New algorithms aiming at automatic analysis of 1H-NMR spectra

MESTRELAB RESEARCH
NMR Solutions

Carlos Cobas [1], Stanislav Sýkora [2]

[1] Mestrelab, Xosé Pasín 6-5C, Santiago de Compostela, Spain 15706; carlos@mestrelab.com
[2] Extra Byte, Via Raffaello Sanzio 22/C, Castano Primo (Mi), Italy 20022; sykora@ebyte.it

DOI: 10.3247/sl2nmr08.005

INTRODUCTION:

Over the years multidimensional NMR has become an essential tool for the structural analysis of molecules but, in practice, the most widely used experiment is still the 1D proton NMR, which, when properly interpreted, gives a wealth of information with minimal acquisition time and sample quantity. However, a detailed analysis of 1H-NMR, especially an automatic one, is often hindered by insufficient resolution (either digital or real), presence of extra lines (solvent and impurities), artifacts due to dead time and shimming, spectral lines overlap, strong coupling effects,

Even though the theory of exact spectral analysis has been worked out over 40 years ago, automatic analysis of 1H-NMR is generally still performed by algorithms based on simple first order rules in which higher order effects (except, perhaps, rudimentary roof effects) are discarded, loosing a great amount of very valuable information.

Here we present some of our recent efforts which aim at an expert system which will overcome the above-mentioned difficulties and get the most out of 1D NMR spectra. If need be, the results obtained in this way can be complemented by information derived from more time consuming NMR techniques, such as 13C and/or multidimensional NMR experiments.

This expert system comprises a number of algorithms for boosting resolution and detecting spectral peaks, complete deconvolution of NMR data sets, quantum-mechanical simulation of spin systems of any size, automatic fitting of experimental spectra, novel ways to sort out the coupling structure between various multiplets, etc. Some of these algorithms extend in a natural way also to 2D spectra. Above all, however, we want to stress two points:

(1) The flowchart-like *modus operandi* of how the verification and elucidation machine must interact with the User in order to accomModate both an attempt at fully automatic solution and a maximum software support to the User in cases (no doubt very frequent) where there are multiple solutions or, apparently, no solution at all.

(2) The fact that verification and elucidation must be carried out in two separate stages: (1) deduction of one or more spin systems compatible with a spectrum and (2) deduction of one or more molecules compatible with a given spin system. The two stages may appear to the User as merged together, but they are actually use quite different software algorithms. For example, the first stage uses spectral simulation and multiplet analysis, while predictions of chemical shifts and coupling constants are useful only in the second stage.

The diagram in **Figure 1 shows the preparatory steps and algorithms** we find necessary before either an automatic or User-aided structure verification or elucidation can be attempted. The blue line on the left indicates the User and the dotted blue arrows pointing towards it indicate the data structures the User should be able to inspect (yellow boxes), even though the execution of this whole Section is automatic. The boxes indicate the following data structures and procedures:

Analog and Digital JC_n : The novel J-Correlator method presented at ENC-2008 [1]

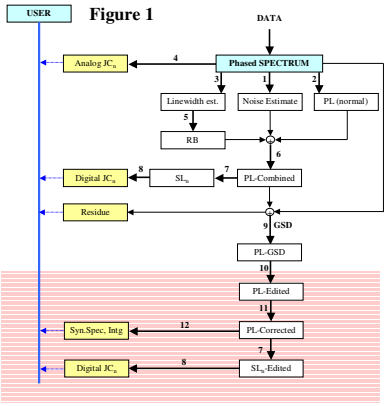
Peaks list (PL): There are several types of them: standard, combined with RB, GSD-refined, edited, and intensity corrected.

Resolution Booster (RB) : The novel peak-localization method presented at ENC-2008 [2]

Splittings list (SL) : List of all splittings between spectral lines which might correspond to a coupling

Global Spectrum Deconvolution : A fitting method to refine parameter of spectral peaks in a Peak List.

Since this Section is executed automatically without the intervention of the User, default Peaks List editing and intensity corrections are used. The principle goal of this stage is a reliable decomposition of a spectrum into a set of spectral lines, removal of baseline and lineshape artifacts and rejection of noise. Thereafter, only the peak list is used for tasks like multiplet coupling structure analysis (**MCSA**), configuration of compatible spin system graphs and fitting of their parameters.



The diagram in **Figure 2 shows the steps leading towards** fitted spin system graphs compatible with the input Peaks List. The dotted blue arrows pointing away from the blue User line indicate interactive User intervention paths (essentially User editing of the input Peaks List and Splittings List and the selection among proposed spin systems). Upon first entry into this interactive loop, evaluation is carried out automatically (with default editing) up to the Spin Systems list and, if there is only one compatible spin system, up to the best-fit peaks list. After every edit, the system automatically re-evaluates all dependencies. Notice that the part marked by the rosy rectangle overlaps with the one in Figure 1 and is repeated here just to make clear the iterative loop. The additional boxes indicate the following data structures and procedures produced by the verification and elucidation machine :

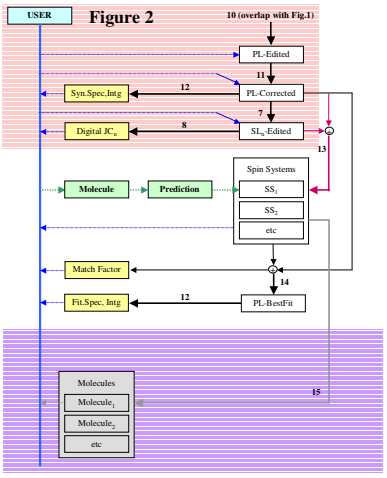
Spin Systems : List of compatible spin systems. In general, more than one spin system may be compatible with a spectrum.

PL Best Fit: List of calculated peaks of the selected spin system after a fit to the edited and intensity corrected input Peak List.

Match Factor : Statistical assessment of the fit quality.

Verification path: the green boxes indicate a typical molecular structure verification path: the User assumes a molecule which uniquely defines a single spin system and, using prediction software, its likely parameters. These are then fitted to the input Peak List and the Match Factor is used to either accept or reject the molecular structure proposal. In the case of rejection, the machine defaults to the complete elucidation loop, while in the case of acceptance the result may be considered definitive (though a match does not guarantee uniqueness).

Note: The assignment of spin systems to a Peaks List (branch 13) is a task employing MSCA tools such as **BIR** (Bayesian Integral Ratios estimator) and the digital J-Correlator. The BIR is a novel algorithm (not yet published) used to find the sets of integer numbers which best fit the experimental multiplet intensities. These, together with the spin system proposal, are also used to carry out intensity corrections of the input Peaks List.



The violet rectangle in **Figure 2** indicates what we have called **Stage 2** in the Introduction. Its task is to *assign molecules to the spin systems* found in the preceding steps which constitute Stage 1. The respective algorithms will rely heavily on prediction software and combinatorial chemistry (structure generators)

Feasibility of structure determination from 1D spectra:

An often asked question is whether we really believe that 1H proton spectra may be sufficient for structure determination. Our answer is that though it will not be possible in general, automatic and/or computer-aided verification and elucidation software can go much further than what has been common so far. In the cases of ^{1}H spectra of clean and relatively small molecules (up to 500 daltons or so) with a number of well separated multiplets, the analysis may actually lead to a single compatible spin system and even a single molecular structure. More often, the User will be presented with two or three possible spin systems and several possibilities of molecular structure which he will have to assess on the basis of his own, unrelated knowledge.

A software of this kind will be therefore appreciated as a valid time-saving tool for chemists working in drug discovery and development and/or natural products analysis. In those cases where the 1D spectrum simply does not contain the desired information, the software will be able to indicate where is the problem and what other information (13C, HSQC, …) is most likely to lead to a rapid solution of the molecular structure puzzle.

References:

[1] Cobas C., Larin N., Iglesias I., Seoane F., Sykora S. **Novel Data Evaluation Algorithms: 1D and 2D Resolution Booster™**, *49th ENC Conference*, Asilomar, CA (USA), March 9-14, 2008.
[2] Cobas C., Monje P., Fraga S., Sykora S., **Novel Data Evaluation Algorithms: J-Correlator™**, *49th ENC Conference*, Asilomar, CA (USA), March 9-14, 2008.

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