The Auto Assignment Algorithm

The Auto Assignment Algorithm combines several software techniques we had developed in recent years as tools for expert tasks such as automatic detection and characterization of spectral peaks, automatic solvent detection, and automatic structure verification (or which the auto-assignment feature is, in its own term, a building block. Real-life spectra always contain a number of artifacts such as noise, baseline distortions, relaxation induced and radiation-damping induced distortions of peak intensities, lineshape distortions due to magnetic field inhomogeneity, lineshape distortions due to unresolved weak long-range couplings, second-order interactions, peaks crowding causing peaks and multiplets to overlap, etc. For these reasons it is impossible to construct any NMR-data evaluation wizard without an extensive usage of statistical methods, allowing for a degree of logical "fuzziness". In our case this is done by applying at every step, to the full depth of the algorithm, a proprietary scoring system approach. The Auto Assignment global flowchart, shown in Fig.1, consists of the following constituent blocks:

1. Basic processing: An NMR-FID is loaded, apodized, transformed, phased and baseline corrected, typically in a transparent, fully unattended way (the process, however, can be predefined by the software user).

2. GSD: The resulting frequency domain 1H spectrum is automatically deconvolved using the sophisticated Global Spectrum Deconvolution algorithm [1] in order to generate a reliable list of peaks and their parameters (position, height, width, curvature, area, etc), even in situations characterized by a strong peaks overlap (Fig.2).

3. AutoClassify: Using another sophisticated fuzzy-logic algorithm [2], each peak in the GSD list is classified according to whether it belongs to the compound or to the solvent, or whether it is an impurity, an artefact, a 1:1 satellite, etc (Fig.3). The algorithm even attempts to pinpoint possible labile peaks.

An important part of this process is also the recognition and of multiplets due to J-couplings and a detailed characterization of their many properties (this results in a multiplets list). Inter-multiplet coupling patterns are also detected and stored internally (the so-called Edited J-C Splittings List) as another tool for the subsequent auto-assignment step. While part of this process can be done without any knowledge of the molecule. This branch of the algorithm points towards automatic spectrum elucidation – a logical future project. In the case discussed here, however, the molecule is presumably known and part of the information about it is used in the AutoClassification process (indicated in the diagram by the arrow from block 4 to block 3).

All this, moreover, is done in an iterative way which confers to the algorithm a capability to ‘backtrack’ and repeat earlier steps with a newly gained or corrected information...

... continued in right column

Introduction

We present an integrated new software solution aimed at automatic assignment of 1H NMR spectra of small molecules. It constitutes an expert system using the principles of fuzzy logic and probabilistic methods which first classifies all the resonances (peaks) in the spectrum and then proceeds to enumerate the most likely assignments of experimental multiplets to a presumed molecular formula and score on them. It uses as inputs the experimental spectrum (or possibly various kinds of spectra) , the suggested molecular formula, and the predicted NMR parameters (shifts and coupling constants) and, as output, it generates the most likely assignment.

The functionality is now available in MestReNova software (www.mestrelab.com).

The Auto Assignment Algorithm

1. Basic processing: An NMR-FID is loaded, apodized, transformed, phased and baseline corrected, typically in a transparent, fully unattended way (the process, however, can be predefined by the software user).

2. GSD: The resulting frequency domain 1H spectrum is automatically deconvolved using the sophisticated Global Spectrum Deconvolution algorithm [1] in order to generate a reliable list of peaks and their parameters (position, height, width, curvature, area, etc), even in situations characterized by a strong peaks overlap (Fig.2).

3. AutoClassify: Using another sophisticated fuzzy-logic algorithm [2], each peak in the GSD list is classified according to whether it belongs to the compound or to the solvent, or whether it is an impurity, an artefact, a 1:1 satellite, etc (Fig.3). The algorithm even attempts to pinpoint possible labile peaks.

An important part of this process is also the recognition and of multiplets due to J-couplings and a detailed characterization of their many properties (this results in a multiplets list). Inter-multiplet coupling patterns are also detected and stored internally (the so-called Edited J-C Splittings List) as another tool for the subsequent auto-assignment step. While part of this process can be done without any knowledge of the molecule. This branch of the algorithm points towards automatic spectrum elucidation – a logical future project. In the case discussed here, however, the molecule is presumably known and part of the information about it is used in the AutoClassification process (indicated in the diagram by the arrow from block 4 to block 3).

All this, moreover, is done in an iterative way which confers to the algorithm a capability to ‘backtrack’ and repeat earlier steps with a newly gained or corrected information...

... continued in right column

Tests with artificial spectra

To validate the performance of the algorithm (as a proof of concept), the system was first tested with the five compounds of Fig. 4 whose NMR spectra were synthesized using the Moinova intrinsic spectrum simulation facility and the ‘theoretical’ NMR spectral parameters generated by NMR predict. The goal was to isolate common practical NMR issues such as peaks overlap, insufficient resolution, etc. Additionally, for each structure, new spectra where calculated in which the chemical shifts were randomly shifted using a normal distribution with SD = 0.25. It was found that of the 69 potential assignments, all were correctly determined by the algorithm except for one labile proton in Quinine.

Tests with experimental spectra

Further tests were conducted with a fully assigned 1H-NMR library consisting of 39 molecules and a total of 355 proton assignments.

In a first test, the performance of the algorithm was evaluated without including into NMRPredict HOSE code DB any of the assigned molecules used in this study. In a second test, some of the molecules where added to the DB in order to get more accurate chemical shift predictions, as well as smaller error bounds. The results obtained when the Auto-Assignment wizard was executed in a fully automatic mode are summarized in the following table:

<table>
<thead>
<tr>
<th>User DB</th>
<th>Correct</th>
<th>Wrong</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>280</td>
<td>75</td>
<td>73</td>
</tr>
<tr>
<td>Yes</td>
<td>295</td>
<td>60</td>
<td>80</td>
</tr>
</tbody>
</table>

Conclusions

The performance of the auto-assignment wizard is quite impressive and usable, especially considering that while the its structure is now well set, its potential is still in development (we might say that it is still learning). Closer inspection to the results showed that any incorrect results are often due to the presence of several assignments with similar probabilities, a situation which will need to be handled.

In this work, only 1D 1H NMR spectra were used but the system is already armed to accept HSQC spectra. Results obtained with a combined 1H & HSQC approach will be covered in a future publication.

References:

[1] Carlos Cobas, Stanislav Sykora, The Bumpy Road towards Automatic Global Spectral Deconvolution (GSD), 50th ENC Conference, Asilomar, CA (USA), March 29-April 4, 2009