Recent Advances in ASV: from Math to NMR

Stan Sykora, Carlos Cobas, Felipe Seoane, et al

Legend:
ASV: Automatic Structure Verification
ASE: Automatic Spectrum Elucidation

www.mestrelab.com

www.ebyte.it
ASV: all the steps, up to the horizon …

✓ **GSD**: Global Spectral Deconvolution
✓ **Scoring systems**: a new mathematical concept
✓ **ASV structure in Mnova**: Tasks & Tests
✓ **Comparing spectra**: NMR data elements of metric sets
✓ **GSD peaks (auto)editing**: the concept
✓ **Solvent recognition**: simple masking & AI approaches
✓ **Labiles, the pesky outcasts**: 3 ways to handle them
✓ **Multiplets**: recognition & characterization
✓ **Counting the nuclei I**: global & regional
✓ **Prediction errors**: definition of prediction regions
✓ **Counting the nuclei II**: prediction regions
✓ **Coupling patterns**: using JC algorithm & predictions
✓ **Assignments**: enumeration and scoring
✓ etc …
GSD: a functional definition

A fully automatic de-convolution of a whole spectrum
GSD history

- Introduction of the idea: SMASH 2007 (Sep, Metrelab User meeting)
- Presentation of first alpha results: 2008 (talks in Italy, UK, China)
- Declared to be fully operative: MMCE 2009 (Feb, a talk)
- Detailed presentation: 50th ENC in 2009 (Mar, poster)
- Significance of GSD for ASV: SMASH 2009 (Sep, User Meeting)
- First official release within Mnova: Autumn 2009
- First major revision: Jan-Feb 2009 (released in March)
- Applications to ASV: coming out now; ENC 2010 (Apr, User meeting)
- Applications to qNMR: in the works (Dr. Peng will present alpha results)
- Lineshape generalization: coming very soon
- ...
GSD algorithm

Four major innovative steps:

1. Derivatives (0th, 1st, 2nd)
2. Special points mark-up
3. Peaks ‘boxing’ (raw GSD)
4. Fast peak fitting
GSD output

An editable Peaks List of all objectively detectable peaks to be used for any subsequent evaluation, including ASV
A novel mathematical concept (as well as a software class) devised to take decisions based on a number of tests, each having its own intrinsic significance.

A situation to which it might apply:

- A committee of 20 votes on a query.
- Each member casts a ‘vote’ or score ranging from –1 (false) to +1 (true).
- Each member, based on his prior record and independently of his vote, is assigned an intrinsic ‘reliability’ or significance (0 to 10).
- The final decision takes into account ALL the (score, significance) pairs.
- The final result is again a pair (score, significance).
- The way the pairs are combined is subject to a set of carefully modeled mathematical constraints (scoring system axioms).

Note: Significance can depend in a predefined way upon the sign of the score (some experts may be better on false scores, others on true scores).
Some scoring system notes

Horizontal axis: Score value (-1 to +1)
Vertical axis: Significance (0 to INF, 1 unreliable, 10 expert)

PASSED and FAILED: Areas of **high-quality** decisions
CAN NOT BE DECIDED: All tests indicate a **decision is impossible**
UNDECIDED: Tests give **contradictory results**
LIKELY: PASSED but with very low significance
UNLIKELY: FAILED but with very low significance
Scoring system: Matlab tests of compliance with *scoring system axioms* (a few examples)

... This is pure applied math; no NMR at all!
ASV structure

Technically, NMR ASV is a software structure which is embodies:

1) A **scoring system** which is its mathematical basis

2) A **pool of tests** (experts, voters)

3) A **set of tasks** (committees) each specialized for some purpose. Each task is composed of a number of tests, each with its own positive and negative significance (tunable parameters)

At present, just one task is implemented: a **generic ASV wizard** using 8 distinct tests. We plan to support a number of tasks drawing on a large pool of tests (hundreds?).
ASV example
(this is an anticipation, just to give you a feeling)

Verifying strychnine molecule against strychnine spectrum
Comparing spectra

Another math concept which is very useful as a tool is that of a distance between spectra

This is because we want to use black-box predictions and just compare the experimental spectrum with a predicted one.

But the concept of a distance could come handy also in other situations (for example finding similar spectra in a data base).
Spectral metric

There were prior proposals of distance-like functionals on pairs of spectra (Bodis, Ross, Pretsch), but they are lacking in some desirable aspects (irregular behavior upon sharp peak overlaps, excessive sensitivity to lineshape, etc.)

We have found a real-valued functional on a pair of spectra which has all the mathematical properties of a metric, avoids the drawbacks of the BRP distance, and is algorithmically compatible with GSD (can be computed directly from the two peak tables).

Distance $d(S_1, S_2)$:

- Is always non-negative: $d(S_1, S_2) \geq 0$
- Is 0 if and only if $S_1 = S_2$ $d(S_1, S_2) \iff S_1 \equiv S_2$
- Is symmetric $d(S_1, S_2) = d(S_2, S_1)$
- Satisfies triangular inequality $d(S_1, S_3) \leq d(S_1, S_2) + d(S_2, S_3)$
Metric tests

\[ d(\text{experimental}, \text{predicted}) > \text{upper\_threshold} \]
\[ d(\text{experimental}, \text{predicted}) < \text{lower\_threshold} \]

Between the thresholds

- FAIRED
- PASSED
- UNDECIDED

One can do it on the whole spectrum or separately on the aliphatic (-0.5 – 2.5), olefinic (2.5 – 5.5) or aromatic (5.5 – 12) regions using two different modes of normalization (one accentuating local quantity, the other local structure).

Hence our first $1 + 2 \times 3 = 7$ tests!
Verifying strychnine molecule against brucine spectrum

<table>
<thead>
<tr>
<th>Individual Tests</th>
<th>Name</th>
<th>Score</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1H Full Spectral Comparison</td>
<td>0.69</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>1H Aromatic Region Comparison</td>
<td>0.37</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>1H Aromatic Region Pattern</td>
<td>-0.59</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>1H Olefinic Region Comparison</td>
<td>0.33</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>1H Olefinic Region Pattern</td>
<td>0.20</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>1H Aliphatic Region Comparison</td>
<td>0.52</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>1H Aliphatic Region Pattern</td>
<td>0.31</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>Test #200</td>
<td>0.31</td>
<td>3</td>
</tr>
</tbody>
</table>

Verifying brucine molecule against strychnine spectrum

<table>
<thead>
<tr>
<th>Individual Tests</th>
<th>Name</th>
<th>Score</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1H Full Spectral Comparison</td>
<td>1.00</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1H Aromatic Region Comparison</td>
<td>1.00</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1H Aromatic Region Pattern</td>
<td>0.76</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>1H Olefinic Region Comparison</td>
<td>1.00</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>1H Olefinic Region Pattern</td>
<td>1.00</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>1H Aliphatic Region Comparison</td>
<td>1.00</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>1H Aliphatic Region Pattern</td>
<td>1.00</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>Test #200</td>
<td>-0.59</td>
<td>4</td>
</tr>
</tbody>
</table>
When comparing experimental and predicted spectra, there are some aspects which need to be addressed before ANY tests can be applied.

Predicted spectra, for example, do not contain any solvent peaks while experimental spectra contain them in unpredictable amounts. The same applies to the reference peaks (such as those of TMS).

Solvent and reference peaks must be located and labeled in the GSD Peaks List of the experimental spectrum. Once labeled, they can be ignored by all subsequent processing algorithms.

Other types of peaks whose recognition and labeling is desirable, such as $^{13}$C satellites, impurities, rotational sidebands, etc.

This often tricky peaks editing process can be done both automatically and/or manually.
Solvent recognition: basics

One way to handle the solvent is zone masking: one labels every peak in more or less ample region around each expected solvent peak. This is quite drastic and often discards a lot of useful info!

A better way is to apply the same know-how an experienced chemist uses (often instinctively) to selectively pick out the solvent peaks even when they are in a crowded region. This means a kind of AI wizard or, more simply, just a pretty clever software.
Solvent recognition: 
masking versus AI wizard
Solvent recognition AI:
let there be a *scoring system* for every line!
Labile protons peaks

This is much the same story, but a more difficult one since there is much less to build upon
Recognition of other special peaks

- $^{13}$C satellites
- Rotation sidebands
- Impurities

This is work-in-progress involving extensive application of scoring systems associated with individual spectral peaks
Multiplets recognition

Work-in-progress

Number of multiplets: 18

<table>
<thead>
<tr>
<th>ppm/ppm</th>
<th>Centroid</th>
<th>nArea</th>
<th>rmsW</th>
<th>Skew</th>
<th>Curt</th>
<th>effPls</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.73 - 11.73</td>
<td>11.73</td>
<td>0.04</td>
<td>0.0</td>
<td>-0.00</td>
<td>+0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>8.09 - 8.06</td>
<td>8.08</td>
<td>3.52</td>
<td>8.1</td>
<td>-0.04</td>
<td>-1.92</td>
<td>5.52</td>
</tr>
<tr>
<td>7.27 - 7.23</td>
<td>7.25</td>
<td>4.04</td>
<td>11.0</td>
<td>-0.00</td>
<td>-1.00</td>
<td>7.65</td>
</tr>
<tr>
<td>7.17 - 7.03</td>
<td>7.13</td>
<td>9.19</td>
<td>30.7</td>
<td>-0.22</td>
<td>-1.52</td>
<td>8.85</td>
</tr>
<tr>
<td>5.95 - 5.91</td>
<td>5.93</td>
<td>1.53</td>
<td>14.5</td>
<td>+0.25</td>
<td>-1.84</td>
<td>4.69</td>
</tr>
<tr>
<td>4.31 - 4.26</td>
<td>4.28</td>
<td>4.96</td>
<td>9.8</td>
<td>+0.08</td>
<td>-1.05</td>
<td>5.38</td>
</tr>
<tr>
<td>4.17 - 3.98</td>
<td>4.06</td>
<td>15.07</td>
<td>50.9</td>
<td>+0.18</td>
<td>-1.39</td>
<td>15.93</td>
</tr>
<tr>
<td>3.89 - 3.83</td>
<td>3.85</td>
<td>4.88</td>
<td>10.6</td>
<td>+0.08</td>
<td>-1.86</td>
<td>2.04</td>
</tr>
<tr>
<td>3.60 - 3.70</td>
<td>3.73</td>
<td>5.07</td>
<td>16.2</td>
<td>+0.42</td>
<td>-0.59</td>
<td>6.00</td>
</tr>
<tr>
<td>3.27 - 3.20</td>
<td>3.25</td>
<td>4.00</td>
<td>13.5</td>
<td>-0.22</td>
<td>-0.94</td>
<td>6.86</td>
</tr>
<tr>
<td>3.10 - 3.09</td>
<td>3.13</td>
<td>7.15</td>
<td>20.4</td>
<td>-0.43</td>
<td>-1.32</td>
<td>3.98</td>
</tr>
<tr>
<td>2.93 - 2.83</td>
<td>2.88</td>
<td>4.81</td>
<td>17.6</td>
<td>-0.06</td>
<td>-0.65</td>
<td>5.25</td>
</tr>
<tr>
<td>2.80 - 2.74</td>
<td>2.76</td>
<td>5.11</td>
<td>14.6</td>
<td>-0.01</td>
<td>-1.97</td>
<td>2.01</td>
</tr>
<tr>
<td>2.69 - 2.60</td>
<td>2.66</td>
<td>5.37</td>
<td>17.9</td>
<td>-0.27</td>
<td>-1.74</td>
<td>4.12</td>
</tr>
<tr>
<td>2.38 - 2.32</td>
<td>2.35</td>
<td>5.10</td>
<td>15.6</td>
<td>+0.06</td>
<td>-1.45</td>
<td>5.20</td>
</tr>
<tr>
<td>1.95 - 1.85</td>
<td>1.90</td>
<td>9.67</td>
<td>11.5</td>
<td>-0.18</td>
<td>+0.08</td>
<td>5.42</td>
</tr>
<tr>
<td>1.49 - 1.44</td>
<td>1.47</td>
<td>5.34</td>
<td>14.6</td>
<td>-0.07</td>
<td>-1.88</td>
<td>4.29</td>
</tr>
<tr>
<td>1.29 - 1.25</td>
<td>1.27</td>
<td>5.15</td>
<td>11.4</td>
<td>-0.02</td>
<td>-1.43</td>
<td>5.23</td>
</tr>
</tbody>
</table>
Scoring on number of nuclei
Scoring on number of nuclei

### Verification Results

<table>
<thead>
<tr>
<th>Molecule</th>
<th>Result</th>
<th>Score</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Passed</td>
<td>0.75</td>
<td>10</td>
</tr>
</tbody>
</table>

**Individual Tests:**

<table>
<thead>
<tr>
<th>Name</th>
<th>Score</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1H Full Spectral Comparison</td>
<td>1.00</td>
<td>2</td>
</tr>
<tr>
<td>1H Aromatic Region Comparison</td>
<td>1.00</td>
<td>2</td>
</tr>
<tr>
<td>1H Aromatic Region Pattern</td>
<td>0.55</td>
<td>3</td>
</tr>
<tr>
<td>1H Olefinic Region Comparison</td>
<td>0.66</td>
<td>3</td>
</tr>
<tr>
<td>1H Olefinic Region Pattern</td>
<td>0.58</td>
<td>3</td>
</tr>
<tr>
<td>1H Aliphatic Region Comparison</td>
<td>1.00</td>
<td>2</td>
</tr>
<tr>
<td>1H Aliphatic Region Pattern</td>
<td>1.00</td>
<td>2</td>
</tr>
<tr>
<td>Test #200</td>
<td>0.13</td>
<td>4</td>
</tr>
</tbody>
</table>
ASV: all the steps, up to the horizon …

- **GSD**: Global Spectral Deconvolution
- **Scoring systems**: a new mathematical concept
- **ASV structure in Mnova**: Tasks & Tests
- **Comparing spectra**: NMR data elements of metric sets
- **GSD peaks (auto)editing**: the concept
- **Solvent recognition**: simple masking & AI approaches
- **Labiles, the pesky outcasts**: 3 ways to handle them
- **Multiplets**: recognition & characterization
- **Counting the nuclei (I)**: global & regional
- **Prediction regions** as defined by prediction error bounds
- **Counting the nuclei (II)** within prediction regions
- **Coupling patterns**: using JC algorithm & predictions
- **Assignments**: enumeration and scoring
- **etc …**

Now, just let the blue line (work-in-progress divide) move down!