Registration and grouping algorithms in protein NMR derived peak lists and their application in protein NMR reference correction

Andrey Smelter¹, Xi Chen², Eric C. Rouchka¹, Hunter N.B. Moseley²,³,⁴

¹Department of Computer Engineering and Computer Science, University of Louisville
²Department of Molecular & Cellular Biochemistry, University of Kentucky
³Markey Cancer Center, University of Kentucky
⁴Resource Center for Stable Isotope Resolved Metabolomics, University of Kentucky

Introduction
Nuclear magnetic resonance spectroscopy of proteins (protein NMR) is a powerful analytical technique for studying structure and dynamics of proteins. Almost all aspects of protein NMR have been accelerated by the development of software tools that enable the analysis of NMR spectral data and its utilization in studying protein structure and dynamics. This includes software for raw NMR processing, spectral visualization, protein resonance assignment, and structure determination. However, full automation of protein NMR data analysis is still a work in progress and data analysis still requires an expert NMR spectroscopist utilizing an array of software tools.

While manual resonance assignment with spectral visualization software is tedious and can take a significant amount of time, a variety of automated and semi-automated assignment programs have been developed to facilitate the protein resonance assignment process, specifically for solution and solid-state NMR. But one of the historical problems that has limited the use of automated and semi-automated protein resonance assignment tools along with other analyses of NMR peak lists is the requirement that users specify uniform match tolerances to perform spin systems grouping and linking or rely on default uniform match tolerance values provided by the tool.

Background
Peak lists derived from both solution and solid-state NMR spectra are commonly used as input for a variety of analyses, especially automated analyses. For these downstream analyses, peak lists must be aligned (registered) to each other and sets of related peaks must be grouped based on common chemical shift dimensions using match tolerance values. However, some subsets of peaks have a smaller variance and can be grouped into spin systems using tighter match tolerance values, while other subsets of peaks have a larger variance in one or all dimensions that require larger match tolerance values for grouping into spin systems for downstream analyses. This is due to the presence of multiple sources of dimension-specific variance in peak positions, which complicates peak grouping and limits the effectiveness of grouping methods that utilize uniform match tolerances. Therefore, we are developing new methods that can detect subsets of peaks with different sources of peak positional variance and group peaks into spin systems based on their specific variance.

Methods

Registration Algorithm
• Has similarities to point pattern matching algorithms.
• Can perform both pairwise- (two different peak lists) and self-registration (single peak list).
• Calculates the best mapping of peaks from the “input” peak list to peaks in the “root” peak list.

Spin System Grouping (Experimental Peak Lists)

Table 1
<table>
<thead>
<tr>
<th>Protein / Peak list</th>
<th>Expected peaks</th>
<th>Observed peaks</th>
<th>Identified spin systems</th>
<th>Missing spin systems</th>
<th>Split spin systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSP / HN(CO)CACB</td>
<td>126 156 57 56 124</td>
<td>12 0 0 0 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ER14 / HN(CO)CACB</td>
<td>191 7 80 67</td>
<td>0 0 0 0 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FGF / HN(CO)CACB</td>
<td>273 303 24 128 139</td>
<td>(112) 13 2 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JR19 / HN(CO)CACB</td>
<td>137 203 36 66 81</td>
<td>(43) 26 8 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NS1 / HN(CO)CACB</td>
<td>235 403 19 116 181</td>
<td>(122) 9 2 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RnaseWT / HN(CO)CACB</td>
<td>235 218 25 88 47</td>
<td>(47) 38 32 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZDOM / HN(CO)CACB</td>
<td>268 240 70 55 56</td>
<td>(56) 1 6 28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GB1 / CANCOCX</td>
<td>268 240 70 55 56</td>
<td>(56) 1 6 28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>268 474 16 55 82</td>
<td>(67) 0 4 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CapGly / NCOCX</td>
<td>410 218 25 88 47</td>
<td>(47) 38 32 5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results

NMR Reference Correction Web Interface

Conclusions
• We have developed a new peak list registration algorithm that is capable of executing in two modes: self-registration and pairwise-registration.
• Self-registration mode allows inferring registration statistics for a single peak list that has multiple peaks per spin system. Pairwise-registration allows alignment of two different peak lists in order to calculate registration statistics.
• Using this self-registration algorithm, we developed a new bottom-up iterative grouping algorithm that can group peaks into spin systems within a single peak list and can handle multiple sources of variance that is present within experimental data sets.
• We have developed automated tools that allowed us to create a very large number of simulated peak lists with a range of positional variance using the entire BMRB and rigorously test the performance and robustness.
• We applied our grouping algorithm to the problem of NMR reference correction for unassigned peak lists (chemical shift values) and created web interface.

Future Directions
Our long-term goal is to develop software tools that will significantly improve the speed and the quality of MAS SSNMR protein resonance assignment. Specifically, we will:
• Finish developing core data structures and algorithms.
• Test, validate and refine computational tools from the standpoint of accuracy, efficiency and robustness.

Acknowledgements
This research is supported by NIH P20GM103436 (KBRIN) and NSF 1252893.