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

Andrev Smelter¹, Xi Chen², Eric C. Rouchka¹, Hunter N.B. Moselev^{2,3,4}

¹Department of Computer Engineering and Computer Science, University of Louisville ²Department of Molecular & Cellular Biochemistry, University of Kentucky ³Markey Cancer Center, University of Kentucky

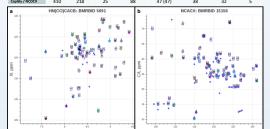
⁴Resource Center for Stable Isotope Resolved Metabolomics, University of Kentucky

Results

Spin System Grouping (Experimental Peak Lists) Table 1. Spin system grouping results for solution NMR derived peak lists using combined registration and

| Protein / Peak list | Expected peaks | Observed peaks | Ungrouped peaks | Expected spin systems | Identified spin systems | Missing spin systems | Overlapped spin systems | Split spin systems |
|--------------------------|----------------|-------------------|-----------------|--------------------------|----------------------------|-------------------------|----------------------------|-----------------------|
| BPTI / HN(CO)CACB | 101 | 134 | 17 | 47 | 54 (30) | 0 | 0 | 2 |
| CSP / HN(CO)CACB | 125 | 145 | 39 | 57 | 53 (32) | 12 | 0 | 0 |
| ER14 / HN(CO)CACB | 194 | 181 | 7 | 93 | 87 (57) | 8 | 2 | 0 |
| FGF / HN(CO)CACB | 273 | 303 | 24 | 128 | 139 (112) | 13 | 2 | 1 |
| JR19 / HN(CO)CACB | 151 | 141 | 7 | 71 | 67 (58) | 4 | 0 | 0 |
| NS1 / HN(CO)CACB | 137 | 203 | 36 | 66 | 81 (43) | 26 | 8 | 2 |
| RnaseC6572S / HN(CO)CACB | 235 | 282 | 16 | 116 | 130 (56) | 18 | 4 | 2 |
| RnaseWT / HN(CO)CACB | 235 | 403 | 19 | 116 | 181 (122) | 9 | 2 | 1 |
| ZDOM / HN(CO)CACB | 134 | 153 | 29 | 67 | 55 (40) | 15 | 3 | 5 |
| ZR18 / HN(CO)CACB | 172 | 163 | 3 | 85 | 80 (52) | 5 | 0 | 0 |

| and grouping a | lgorithm. | | | | | | | |
|---------------------|----------------|-------------------|-----------------|-----------------------|----------------------------|-------------------------|----------------------------|-------|
| Protein / Peak list | Expected peaks | Observed peaks | Ungrouped peaks | Expected spin systems | Identified spin systems | Missing spin systems | Overlapped spin systems | Split |
| GB1 / CANCOCX | 268 | 240 | 70 | 55 | 56 (56) | 1 | 6 | 2 |
| GB1 / NCACX | 268 | 463 | 62 | 55 | 65 (65) | 0 | 0 | 1 |
| GB1 / NCOCX | 268 | 474 | 16 | 55 | 82 (67) | 0 | 4 | 1 |
| DsbB / NCACX | 940 | 215 | 43 | 175 | 47 (47) | 126 | 14 | 1 |
| / | 440 | | 4.0 | 00 | E0 (E0) | 2.2 | 25 | |



Spin System Grouping (Simulated Peak Lists)

| Table 3. Simulated HN(CO)CACB peak lists. | | | , | |
|--|---|---|---|--|
| Number of variance sources | Minimum standard deviation values | Maximum standard deviation values | Total number of simulated peak lists | |
| Single source of variance in all dimensions | H: 0.001 C: 0.01 N: 0.01 | H: 0.050 C: 0.50 N: 0.50 | 127,450 | |
| Two sources of variance in all dimensions | H: 0.001, 0.005 C: 0.01, 0.05 N: 0.01, 0.05 | H: 0.010, 0.050 C: 0.10, 0.50 N: 0.10, 0.50 | 25,490 | |
| Two sources of variance in N dimension, single source of variance in C and H dimensions | H: 0.001 C: 0.01 N: 0.01, 0.05 | H: 0.010 C: 0.10 N: 0.10, 0.50 | 25,490 | |

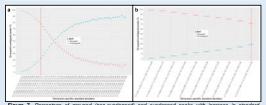


Figure 7. Percentage of grouped (non-overlapped) and overlapped peaks with increase in standard deviation values of peak dimensions: a) single source of variance in all dimensions; b) two sources of variance in all dimensions (20% of peaks have five times larger variance than the remaining 80% of peaks).

Results (continued)

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
- 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

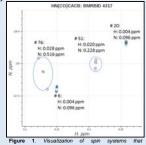
Nuclear magnetic resonance spectroscopy of proteins (protein NMR) is a powerful analytical technique for studying structure and dynamics of proteins. Almost all aspects of protein NMR have been accelerated by the development of software tools that enable the analysis of NMR spectral data and its utilization 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 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.



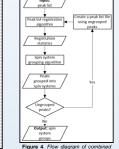
nce within HN(CO)CACB peak list



Registration Algorithm

Methods





Has similarities to

point pattern match

Can perform both

pairwise- (two dif-

ferent peak lists)

and self-registration

Calculates the best

mapping of peaks from the "input"

peak list to peaks in

the "root" peak list.

(single peak list).

algorithms.

- Based on the widely-used density-based clustering algorithm DBSCAN, which can detect clusters of varying size and shape.
- Combines both the self-registration algorithm and grouping algorithm to derive spin system clusters using multiple variancebased match tolerances in an iterative algorithm.

Peak List Simulation Algorithm



- Can simulate peak lists using assigned chemical shift values deposited in RMRR entries
- Uses the nmrstarlib package functionality to access assigned chemical shift values for H. C and N resonances and has an ability to add varying amount of noise.