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

Andrey Smelter[1], Xi Chen[2], Eric C. Rouchka[1], Hunter N.B. Moseley[2,3,4]

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

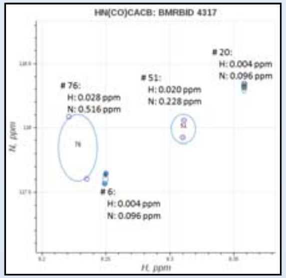


Figure 1. *Visualization of spin systems that demonstrates the presence of multiple sources of variance within HN(CO)CACB peak list.*

♦ Methods

Registration Algorithm

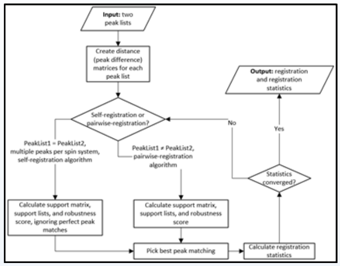

Input: two peak lists
Create distance (peak difference) matrices for each peak list
Self-registration or pairwise-registration?
PeakList1 = PeakList2, multiple peaks per spin system, self-registration algorithm
PeakList1 ≠ PeakList2, pairwise-registration algorithm
Calculate support matrix, support lists, and robustness score, ignoring perfect peak matches
Calculate support matrix, support lists, and robustness score
Pick best peak matching
Calculate registration statistics
Statistics converged?
No
Yes
Output: registration and registration statistics

Figure 2. *Flow diagram of registration algorithm.*

- Has similarities to point pattern match 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.

Grouping Algorithm

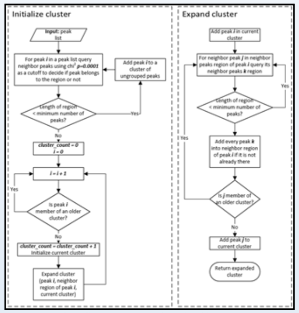

Initialize cluster
Input: peak list
For peak i in a peak list query neighbor peaks using chi² p=0.0001 as a cutoff to decide if peak belongs to the region or not
Add peak i to a cluster of ungrouped peaks
Length of region < minimum number of peaks?
Yes
No
cluster_count = 0
i = 0
i = i + 1
Is peak i member of an older cluster?
Yes
No
cluster_count = cluster_count + 1
Initialize current cluster
Expand cluster (peak i, neighbor region of peak i, current cluster)
Expand cluster
Add peak i in current cluster
For neighbor peak j in neighbor peaks region of peak i query its neighbor peaks k region
Length of region < minimum number of peaks?
Yes
Add every peak k into neighbor region of peak i if it is not already there
Is j member of an older cluster?
Yes
No
Add peak j to current cluster
Return expanded cluster

Figure 3. *Flow diagram of grouping algorithm.*

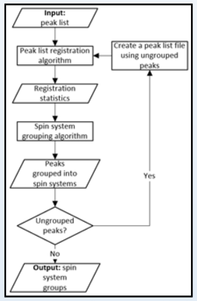
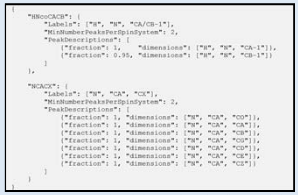

Input: peak list
Peak list registration algorithm
Create a peak list file using ungrouped peaks
Registration statistics
Spin system grouping algorithm
Peaks grouped into spin systems
Ungrouped peaks?
Yes
No
Output: spin system groups

Figure 4. *Flow diagram of combined algorithm.*

- 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 variance-based match tolerances in an iterative algorithm.

Peak List Simulation Algorithm



```
{
    "HNcoCACB": {
        "Labels": ["H", "N", "CA/CB-1"],
        "MinNumberPeaksPerSpinSystem": 2,
        "PeakDescriptions": [
            {"fraction": 1,    "dimensions": ["H", "N", "CA-1"]},
            {"fraction": 0.95, "dimensions": ["H", "N", "CB-1"]}
        ]
    },

    "NCACX": {
        "Labels": ["N", "CA", "CX"],
        "MinNumberPeaksPerSpinSystem": 2,
        "PeakDescriptions": [
            {"fraction": 1, "dimensions": ["N", "CA", "CO"]},
            {"fraction": 1, "dimensions": ["N", "CA", "CA"]},
            {"fraction": 1, "dimensions": ["N", "CA", "CB"]},
            {"fraction": 1, "dimensions": ["N", "CA", "CG"]},
            {"fraction": 1, "dimensions": ["N", "CA", "CD"]},
            {"fraction": 1, "dimensions": ["N", "CA", "CE"]},
            {"fraction": 1, "dimensions": ["N", "CA", "CZ"]}
        ]
    }
}
```

Figure 5. *Spectrum description configuration file of peak list simulation algorithm.*

- Can simulate peak lists using assigned chemical shift values deposited in BMRB 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.

♦ Results

Spin System Grouping (Experimental Peak Lists)

Table 1. *Spin system grouping results for solution NMR derived peak lists using combined registration and grouping algorithm.*

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

Table 2. *Spin system grouping results for solid-state NMR derived peak lists using combined registration and grouping algorithm.*

Protein / Peak list	Expected peaks	Observed peaks	Ungrouped peaks	Expected spin systems	Identified spin systems	Missing spin systems	Overlapped spin systems	Split spin systems
GB1 / CANCOCX	268	240	70	55	56 (56)	1	6	28
GB1 / NCACX	268	463	62	55	65 (65)	0	0	19
GB1 / NCOCX	268	474	16	55	82 (67)	0	4	10
DsbB / NCACX	940	215	43	175	47 (47)	126	14	1
CapGly / NCACX	410	515	16	88	50 (50)	33	25	0
CapGly / NCOCX	410	218	25	88	47 (47)	38	32	5

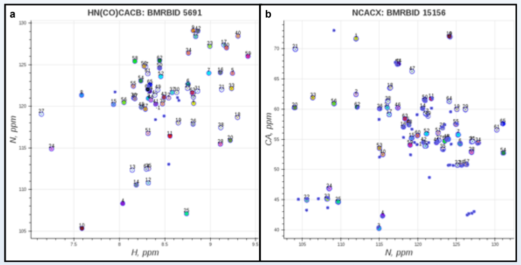


Figure 6. *Visualization of spin systems:* **a)** *example of best spin system clustering for 30S ribosomal protein S28E from Pyrococcus horikoshii protein;* **b)** *example of best spin system clustering for GB1 protein.*

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

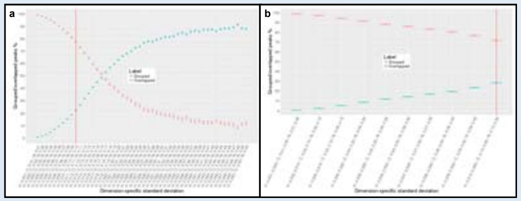
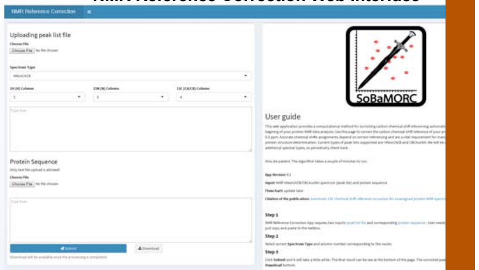


Figure 7. *Percentage of grouped (non-overlapped) and overlapped peaks with increase in standard deviation values of peak dimensions:* **a)** *single source of variance in all dimensions;* **b)** *two sources of variance in all dimensions (20% of peaks have five times larger variance than the remaining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

♦ Results (continued)

NMR Reference Correction Web Interface



♦ Conclusions

- We have developed a new peak list registration algorithr capable of executing in two modes: self-registrati pairwise-registration.
  - Self-registration mode allows inferring registration s for a single peak list that has multiple peaks per spin s
  - Pairwise-registration allows alignment of two differe lists in order to calculate registration statistics.
- Using this self-registration algorithm, we developed bottom-up iterative grouping algorithm that can group pe spin systems within a single peak list and can handle sources of variance that is present within experimen sets.
- We have developed automated tools that allowed us to very large number of simulated peak lists with a ra positional variance using the entire BMRB and rigorou the performance and robustness.
- We applied our grouping algorithm to the problem reference correction for unassigned peak lists (chemi values) and created web interface.

♦ Future Directions

Our long-term goal is to develop software tools tl significantly improve the speed and the quality of MAS protein resonance assignment. Specifically, we will:

- Finish developing core data structures and algorithms.
- Test, validate and refine computational tools from the sta of accuracy, efficiency and robustness.

♦ Acknowledgements

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