AUTOMATED PROTEIN NMR DATA ANALYSIS AND ITS APPLICATION TO α-SYNUCLEIN FIBRILS

BY

JOSEPH MICHAEL COURTNEY

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Chemistry
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Professor Chad Rienstra, Chair
Professor Martin Gruebele
Professor Sharon Hammes-Schiffer
Professor Eric Oldfield

ABSTRACT

In principle, nuclear magnetic resonance (NMR) spectroscopy provides structural and conformational information with sub-Angstrom precision and the ability to measure dynamics with timescales ranging from femtoseconds to years, all with atomic specificity. However, due to the relatively low sensitivity of NMR, fundamental limits on spectral resolution, and the complexity of the quantum mechanical phenomena NMR exploits, that wealth of information often remains out of reach. The highly varied presentation of molecular information in NMR spectra and the difficulty of numerical simulation of non-trivial systems has lead the majority of data analysis to be performed by trained experts, and because of its time-intensive nature, that analysis is rarely replicated by a third party or validated in an objective manner.

In this dissertation I report my efforts to automate NMR data analysis in an objective and replicable manner and to provide tools for validation of resulting three-dimensional structures by direct comparison to raw spectral data. The first method, COMPASS, attempts to extract as much information as possible from a single ^{13}C-^{13}C two-dimensional spectrum for the determination of protein structure and successfully identified the true structure of 15 test proteins. The second method, GPS, predicts features of data that would be expected given a set of chemical shift assignments and possibly a three-dimensional structure and uses the the presence or absence of those features in experimental spectra to refine or validate a given structure.

I then report my application of these computational methods to the problems of refining an α-synuclein fibril structure with proton-detected NMR data, the analysis and characterization of a pair of interrelated α-synuclein fibril strains with distinct pathological properties, and to the general question of fibril polymorphism, a phenomenon that presents a substantial challenge to forming consistent conclusions about fibril properties and interactions across samples and research groups.

To my wife Elizabeth Courtney,

who braved the frozen northern wasteland and countless late dinners to indulge my curiosity.

To my son Benjamin Wolfgang “Wolfy” Courtney,

whose snores, squeaks, and coos accompanied the writing of most of this dissertation.

TABLE OF CONTENTS

INTRODUCTION

Nuclear magnetic resonance (NMR) spectroscopy is the measurement of miniscule currents in a resonant coil induced by an oscillating bulk magnetic moment caused by the collective transition of nuclei in a sample from a parallel alignment to an antiparallel alignment with a high strength external magnetic field. The spin states of the nuclei evolve according to a sequence of Hamiltonians with contributions from interactions within the sample, including the Zeeman effect, quadrupolar couplings, dipolar couplings, chemical shielding, and J-couplings, and external manipulations in the form of radio frequency pulses, magnetic field gradients, and mechanical rotation. By careful design of external manipulations, the measured signal can be made to encode different combinations of the effects of the NMR Hamiltonian that encode enormous amounts of information about the sample. However, the complexity of the interactions involved and the bulk nature of the measurement compound inherent difficulties such as the basic insensitivity of NMR and the extensive sampling necessary to measure the high complexity spectra to sufficient resolution.

Due to the complexity of the data, the necessary compromises made during data collection, and the difficulty of modelling protein-scale quantum interactions involving all possible conformational and dynamical states, the processing and analysis of NMR data on samples more complicated than small organic molecules are primarily done manually. Manual analysis is problematic in many ways including its time-consuming nature, the requirement of extensive training, and most importantly, the possibility of human bias and error affecting the final results of spectral analysis. As such, the need for computational and automated methods for analyzing and data and scrutinizing the resulting models is paramount.

This dissertation details my work to automate analysis and objectively evaluate the resulting conclusions of that analysis for consistency with experimental data. In chapter 1, I describe my efforts to objectively assess the consistency of protein structures with experimental data. In chapter 2, I detail the proper propagation of uncertainty from raw experimental data through spectral reconstruction methods and analysis to the point of spectral summarization as a set of peaks. In chapter 3, I provide a new method for fitting dipolar coupling trajectories that avoids the imposition of structural and dynamical models. Chapter 4 presents the concept of Global Peak Scoring (GPS), a flexible, general purpose method for enumerating the signals that should be present in a correlation spectrum of a protein, assuming a structural model and model of the coherence transfer dynamics; this approach provides a method of checking analysis for self-consistency. Chapter 5 describes my application of the GPS method to the refinement of a structure of pathogenic α-synuclein fibrils. In chapter 6, I explain my analysis of a pair of protein fibrils using a primarily traditional, manual method of analysis and a set of conclusions that will benefit from the application of self-consistency checks. Chapter 7 describes a forward-looking plan for a method to check the consistency of a set of conclusions with the complete dataset used to make them.

CHAPTER 1: Experimental Protein Structure Verification by Scoring with a Single, Unassigned NMR Spectrum

Introduction

NMR is a powerful technique for studying protein structure and dynamics in near-native conditions. Substantial progress has been made in the solution of high-resolution protein structures by solid-state NMR (SSNMR) in the last decade. Structures which were previously inaccessible by solution NMR and X-ray crystallography, such as fibrils of the HET-s protein and amyloid-β, have been solved at atomic detail, offering insight into important biomedical problems. (Wasmer et al. 2008, Lu et al. 2013) SSNMR approaches to solving structures of membrane proteins also have several notable successes. (Shahid et al. 2012, Wang et al. 2013, Park et al. 2012)

However, NMR methods, and SSNMR in particular, still require extensive sample preparation, data collection, and interpretation efforts. Typically, tens of milligrams of ^{13}C,^{15}N-labeled protein and several weeks of instrument time are required, in order to collect a half dozen or more 3D data sets necessary for the resonance assignments. Additional samples with sparse ^{13}C labeling and weeks of instrument time are needed to obtain a sufficient number of inter-residue distances to determine the fold uniquely. (Comellas et al. 2013) Methods are in development to shorten the lengthy process of data collection, including non-uniform sampling (NUS) (Paramasivam et al. 2012, Hyberts et al. 2010, Sun et al. 2012), proton detection with fast magic-angle spinning (MAS) (Knight et al. 2011, Zhou et al. 2012, Barbet-Massin et al. 2014) and combinations of these two approaches. (Linser et al. 2014). Dynamic nuclear polarization is also a very promising method for accelerating data collection times, yet is usually not compatible with

conditions that yield high-resolution spectra. (Maly et al. 2008, Wang et al. 2013, Renault et al. 2012)

In addition to challenges associated with data collection, the assignment and interpretation of spectra to yield a structure remain major bottlenecks and can take months of manual data analysis. Although methods are now available to automate the assignment process (Moseley et al. 2010, Güntert 2009, Guerry & Hermann 2011, Schmidt et al. 2013), these approaches still require complete sets of 3D data and extensive manual intervention. Once resonance assignments are available, methods such as CS-ROSETTA (Shen et al. 2008) and CHESHIRE (Cavalli et al. 2007, Robustelli et al. 2010) are available to leverage the chemical shift data for structure determination. These approaches have been highly successful; yet still require complete sets of site-specific resonance assignments. Therefore, there remains a compelling need for alternative methods that are faster and more cost-effective, requiring less sample, instrument time, and analysis. Combining NMR with advances in protein structure prediction (both homology modeling and *ab initio* methods) offers a potential increase in efficiency, (Simons et al. 1997, Eswar et al. 2002, Moult et al. 2014) This approach requires validation by comparing predicted NMR observables from the models with empirical or experimental data. In all prior methods, this has been done using sequence-specific resonance assignments.

Here we present a method, called Comparative, Objective Measurement of Protein Architectures by Scoring Shifts (COMPASS) that aims to extract structural information from NMR spectra by fully leveraging a limited amount of experimental data—one 2D ^{13}C-^{13}C spectrum—to accurately distinguish the correct protein fold from a set of proposed models. This avoids the lengthy structure determination process and requires no manual analysis of spectra.

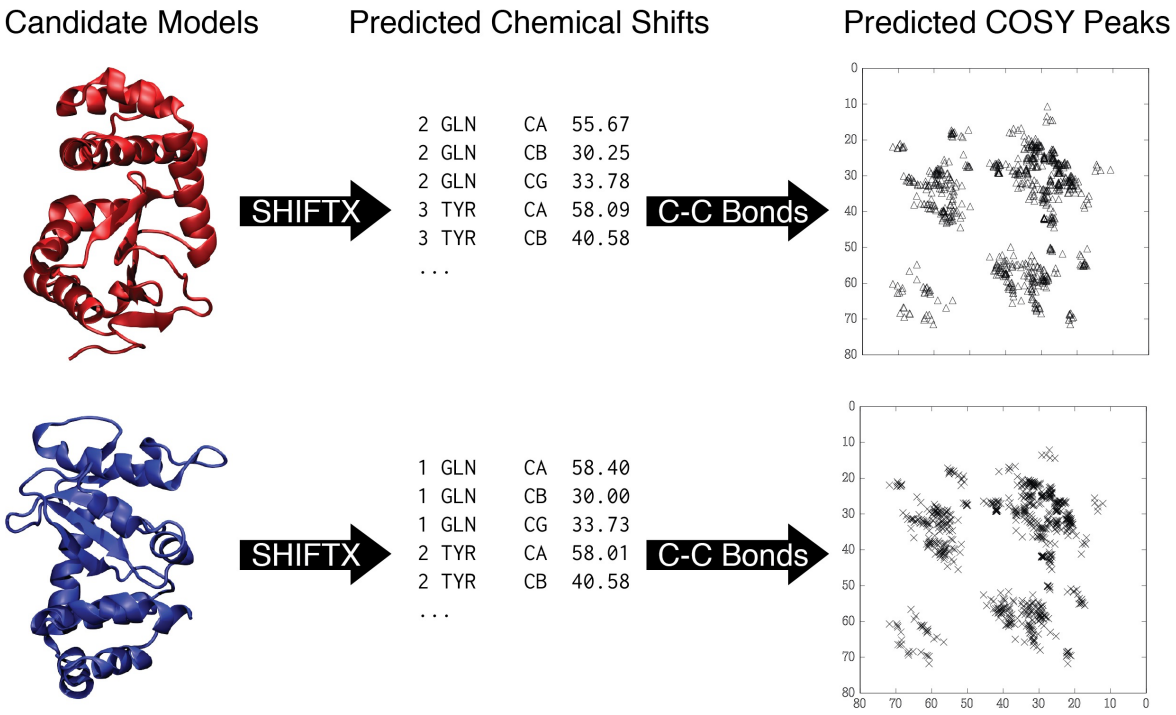

Figure 1.1 – Prediction of 13C-13C correlation spectra from protein models with SHIFTX. The predicted chemical shifts are paired using a python function that enumerates all directly bonded carbon pairs in the structure and the corresponding chemical shifts are stored in a list without any assignment information.

COMPASS solely employs the numerical comparison of predicted spectra from structural models, produced by various methods (e.g., homology modeling, molecular dynamics, *ab initio* quantum chemistry), with a single, unassigned 2D ^{13}C-^{13}C NMR spectrum, utilizing the dependence of chemical shifts upon protein conformation.

COMPASS leverages the accuracy of ^{13}C chemical shift prediction methods, and in this study we utilize SHIFTX2 (Han et al. 2011). For each protein, we collect a ^{13}C-^{13}C homonuclear correlation spectrum under conditions of scalar or dipolar mixing that yield exclusively one-bond correlations throughout the entire aliphatic region. (Chen et al. 2006, Hohwy et al. 1999) Cross peaks in this spectrum are enumerated and filtered according to a simple heuristic to generate a list of unassigned peaks. Meanwhile, a series of models are generated from the amino acid sequence using either homology or *ab initio* methods, and the ^{13}C chemical shifts are predicted for each model by SHIFTX2. Due to the simplicity and predictability of single-bond homonuclear

correlation spectra, the hypothetical cross peaks that would result from each model can be predicted (Fig. 1.1). Then, using a scoring method based on the modified Hausdorff distance, (Dubuisson et al. 1994) the models can be ranked according to their consistency with the experimental peak list. In the large majority of cases, the best model identified is consistent with the experimentally solved structure.

Results and Discussion

To test COMPASS, we selected sixteen proteins ranging in molecular mass from 6.6 to 33.6 kDa. For all selected proteins, high-quality structures of the monomeric form in the absence of any perturbing ligands are available in the Protein Data Bank. (Bernstein et al. 1977) Two-dimensional one-bond ^{13}C-^{13}C correlation spectra under solid-state conditions (MAS) were collected for four of these proteins—GB1, ubiquitin, DsbA, and the extracellular domain of human tissue factor (TF). For GB1, ubiquitin, and DsbA, CTUC-COSY spectra were collected. For TF, we collected an SPC5 spectrum with a short mixing time to observe only one-bond transfers. (Hohwy et al. 1999) Other pulse sequences that generate one-bond correlations could also be employed.

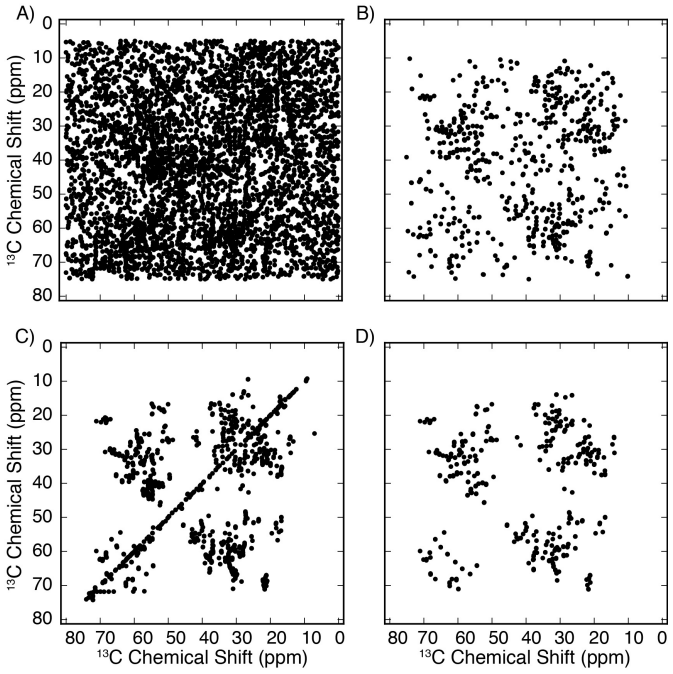

Figure 1.2 - Peak filtering procedure. (A) Peaks automatically picked in the Sparky analysis program with a noise floor set at twice the root mean squared (RMS) noise level. (B) The same peaks after being filtered to exclude points near the diagonal and peaks without corresponding peaks opposite the diagonal. (C) Peaks automatically picked with a noise floor set at six times the RMS noise level. (D) The same data as (C), but filtered as in (B).

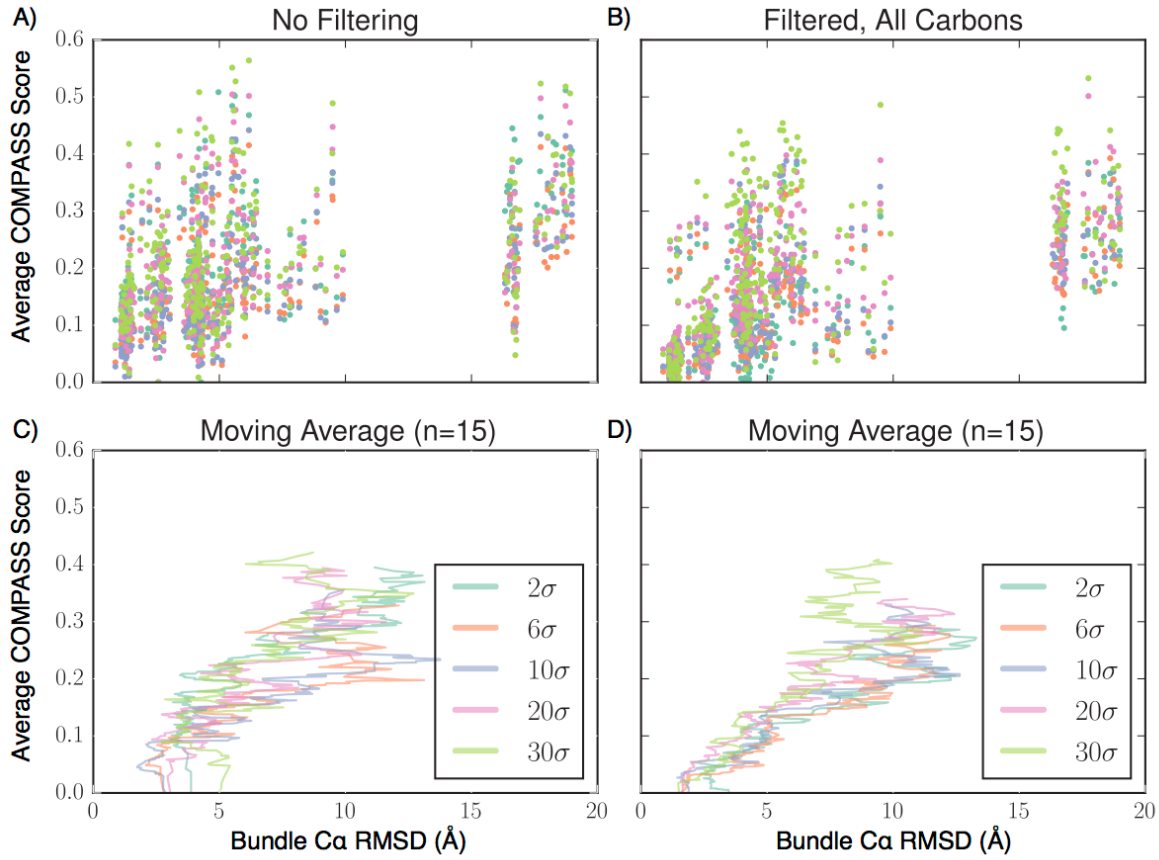

Figure 1.3 - Effect of differing noise floors on convergence of COMPASS scoring on DsbA. Parts (A) and (B) show COMPASS score vs. C✔ RMSD for the peaks picked without filtering (A) and with filtering (B). Each point set is colored according to the noie floor, given in multiples of σ, the RMS noise in the data. In parts (C) and (D) the vertical axes are the moving average (n=15) of COMPASS scores of models ranked by individual COMPASS scores and the horizontal axes are the bundle C✔ RMSD (compared to 1FVK) for each model set used in the COMPASS score moving average. (C) is for peaks without filtering and (D) is for filtered peaks.

Automated Peak Filtering

Peaks were picked using the automated peak picking function of the Sparky NMR data analysis program. (Goddard et al. 1999) A range of noise floors was tested and an optimal minimum signal-to-noise ratio (SNR) of 6 was chosen on the basis of testing shown in figure 1.2. Peaks were then filtered to retain only those in the aliphatic region (0-80 ppm), at least 0.5 ppm away from the diagonal. The lists were then further filtered to retain only those peaks that were observed on both sides of the diagonal within a cutoff of 0.3 ppm (Fig. 1.2b, 1.2d). This automated

peak picking and filtering heuristic contributes significantly to the noise tolerance of COMPASS, as observed by the exclusion of the majority of the noise peaks even in a spectrum picked with a noise floor of twice the RMS noise (Fig. 1.2b).

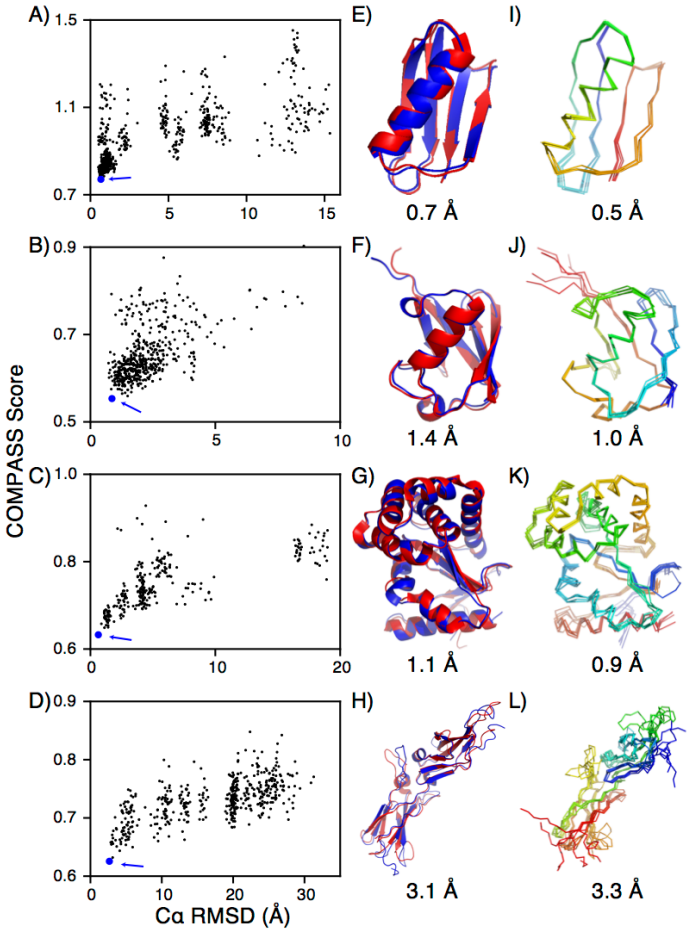

Figure 1.4 - COMPASS results for four proteins with unassigned NMR data. COMPASS Score vs. C✔ RMSD from the reference structure for (A) GB1 (1FVK), (B) Ubiquitin (1UBQ), (C) DsbA (1FVK), and (D) TF (1BOY). The structure with the lowest COMPASS score is shown in blue and indicated with an arrow. (E-H) The lowest dMHD structure (blue) overlaid with the reference structure (red). The C✔ RMSD is noted. (I-L) The five lowest dMHD structures aligned and overlaid. The average pairwise C✔ RMSD is noted.

Evaluation of COMPASS Score

To test the behavior of the COMPASS score on models of differing accuracy, we investigated the relationship between the scores of a group of models and their Cα RMSDs measured against the reference structure deposited in the Protein Data Bank. Figure 1.4 shows plots of the COMPASS score vs. Cα RMSD for the four proteins with peak lists obtained directly from 2D spectra. For all four examples, models with lowest scores have low RMSDs. The obverse, however, is not always true. As can be seen, especially for GB1 (Fig. 1.4A), many models with RMSDs below 2 Å have scores greater than or equal to those models with RMSDs > 10 Å. This phenomenon occurs because the scores depend not only on the Cα-Cβ correlations, which report most strongly on secondary structure, but also on cross peaks involving side chain carbons, which report more

strongly on the local environment. (Han et al. 2011) Therefore models with the correct side chain conformations will agree best with the NMR data (i.e., exhibit the lowest scores). This behavior gives the COMPASS score a conservative character in that it rejects some models that have good coarse grain structure but incorrect side chain packing, while uniformly rejecting models with incorrect folds. Consistent with the score's sensitivity to side chain conformation, there is a decreased correlation between the score and RMSD at higher RMSD values, since models with extremely different backbone structure but energetically optimized side chains are very unlikely to have conformations that would produce similar side chain ^{13}C chemical shifts.

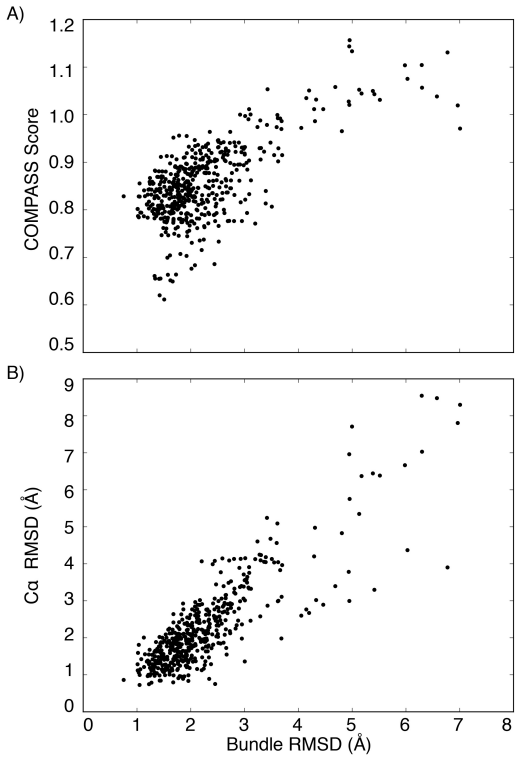

Figure 1.5 - Ordered bundle RMSD. Models are scored and ordered by the COMPASS scores. The bundle RMSD is the average RMSD of the four models with COMPASS scores closest to its own. (A) COMPASS score vs. bundle RMSD showing the "funneling" towards the origin indicative of a data set containing a correct consensus structure. (B) The bundle RMSD is highly correlated with the C✔ RMSD to the correct structure, which enables its use as a surrogate when the true structure is unknown.

Overlays of the reference structure (red) with the model with the lowest score (blue) for each protein are shown in Fig. 1.4e-h. For all tested proteins, the bundle RMSD acts as a good surrogate for the actual RMSD from the true structure. When the bundle of five lowest-score structures had an acceptably small average pairwise RMSD, the consensus structure also had a low RMSD with respect to the reference structure (Fig. 1.5).

To test the performance of COMPASS on a wider range of structures, we chose an additional 11 proteins with known structure and complete ^{13}C chemical shift assignments from the Biological Magnetic Resonance Data Bank (BMRB). (Ulrich et al. 2008) In lieu of raw spectra,

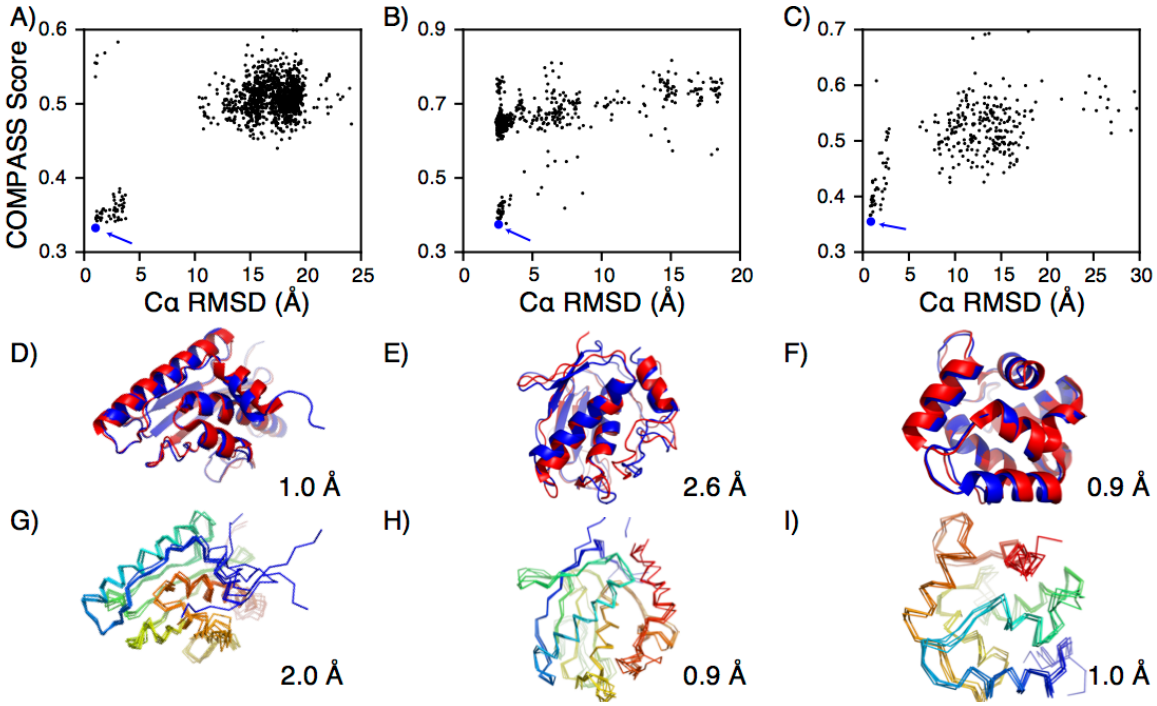

Figure 1.6 - Additional COMPASS results for synthetic peak lists constructed from BMRB-depositied chemical shifts. COMPASS score vs. C✔ RMSD from the reference structure for (A) Basic fibroblast growth factor (1BFG), (B) Sterol carrier protein 2 (1C44), and (C) Integrin alpha-L (1XUO).

we reconstructed peak lists from the known assignments using the same algorithm applied to predicting model peak lists. Although the sequence-specific assignments were available for these cases, the assignment information was not carried forward in the calculation.

The COMPASS score performed similarly well for most proteins in the synthetic data set (Figs. 1.6-1.10). However, for the protein StR65, none of the models predicted by MODELLER had an RMSD below 10 Å. For this data set, the COMPASS score exhibits the desirable quality that

the five structures that agree most closely with the experimental data have an average pairwise RMSD of over 22.4 Å, providing an unambiguous indication that a consensus structure does not exist in the model set (0. 11b). As expected, if the set of models supplied to COMPASS does not contain any models that are consistent with the experimental data, a consensus structure cannot be identified.

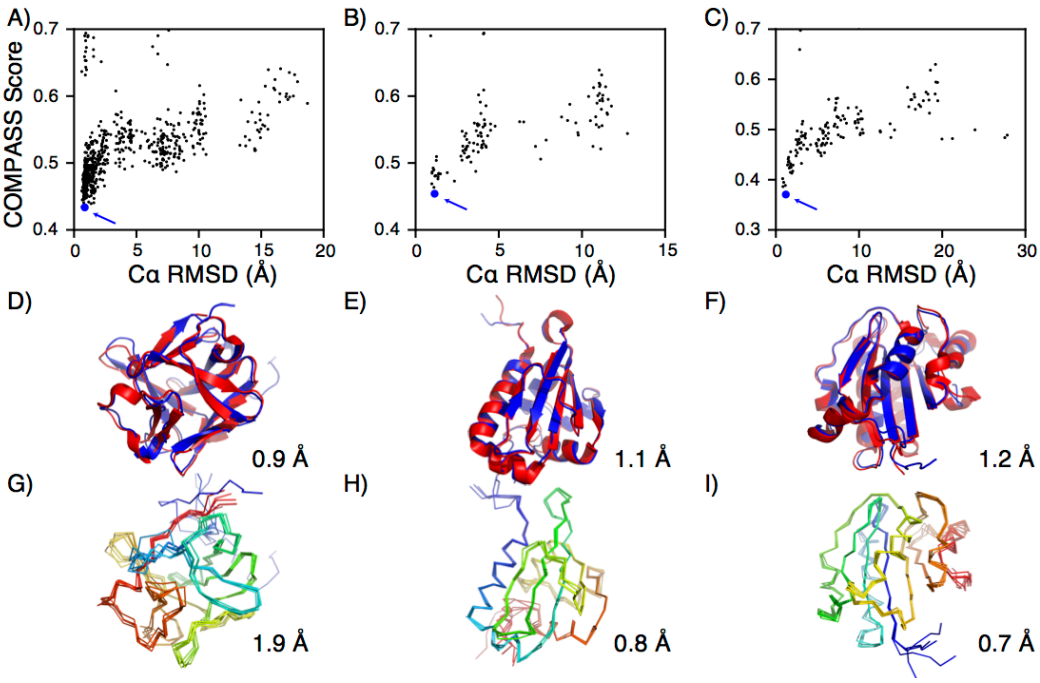

Figure 1.7 - Additional COMPASS results for synthetic peak lists constructed from BMRB-deposited chemical shifts. COMPASS score vs. C✔ RMSD from the reference structure for (A) Ufm1-conjugating enzyme 1 (2Z6O), (B) Macrophage metalloelastase (2KRJ), (C) Alpha-parvalbumin (1RWY). The structure with the lowest COMPASS score is shown in blue and indicated with an arrow. (D-F) The structure with the lowest COMPASS score (blue) overlaid with the reference structure (red). The C✔ RMSD is noted. (G-I) The overlay of five structures from each calculation with the lowest COMPASS scores. The average pairwise C✔ RMSD is noted.

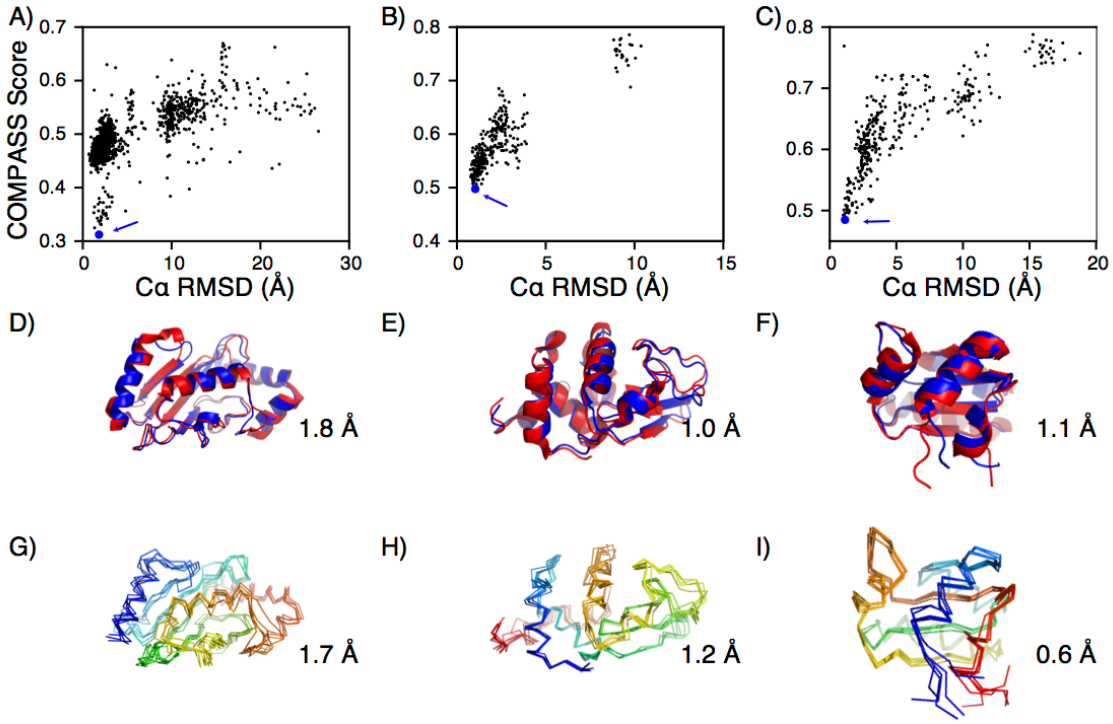

Figure 1.8 - Additional COMPASS results for synthetic peak lists constructed from BMRB-depositied chemical shifts. COMPASS score vs. C✔ RMSD from the reference structure for (A) Ubiquitin-conjugating enzyme 1 (1FZY) (B) Lysozyme C (1IWT) (C) 50S ribosomal protein L30E (1GO1).

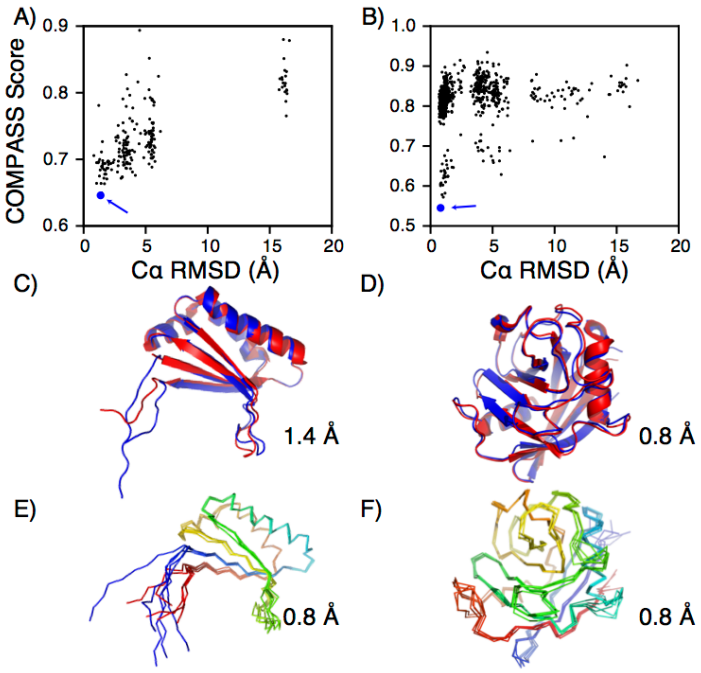

Figure 1.9 - Additional COMPASS results for synthetic peak lists constructed from BMRB-depositied chemical shifts. COMPASS score vs. C✔ RMSD from the reference structure for (1GO1) (A) Protein At3g7210 (1Q4R) (B) Cyclophilin A (2CPL).

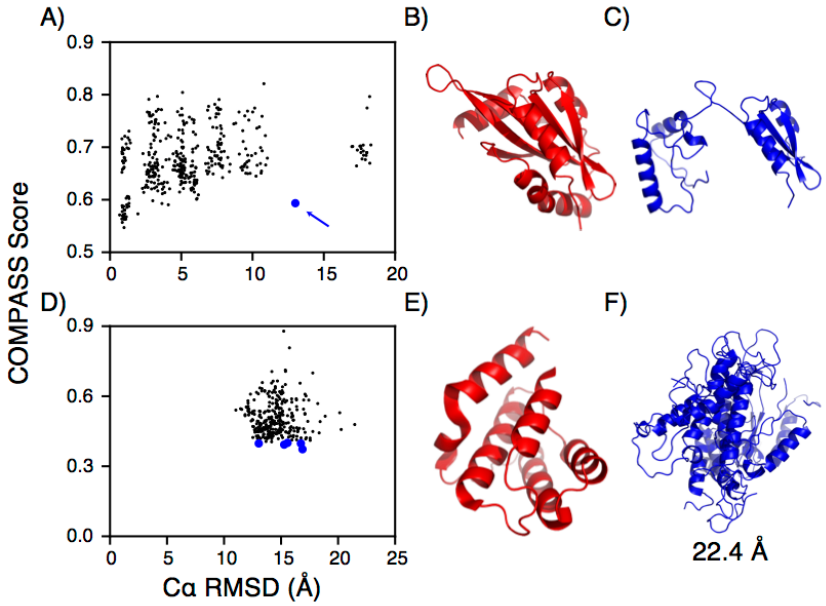

Figure 1.10 - The behavior of the COMPASS scoring method when applied to incorrect models. (A-C) Coactosin-like protein (A) COMPASS score vs. C✔ RMSD from (1T3Y). Point with anomalously low score is blue and noted with an arrow. (B) Structure from 1T3Y. (C) Structure of outlier model showing split structure. (D-F) NorthEast Structural Genomics consortium target STR65 (D) COMPASS score vs. C✔ RMSD from (2ES9). Points with 5 lowest COMPASS scores are denoted by large blue dots. (E) Structure from 2ES9. (F) Aligned overlay of five lowest COMPASS score structures. C✔ RMSD is noted.

In one case, a model with a low score but a high RMSD was observed. In this calculation on coactosin-like protein, a single model was generated with a Cα RMSD of 13 Å but had a COMPASS score comparable to much better models (Fig. 1.7a). Upon manual inspection of the outlying model, it is clear that the majority of the secondary and tertiary structure elements are correct, but the model corresponds to a protein with two domains dissociated from each other, tethered by an unstructured loop. While this outlier did not perform as expected, its score is still well above that of the consensus, which agrees with the reference structure to within an RMSD of 0.72 Å. Manual inspection or the application of structure validation programs would easily identify this model as incorrect and it could be removed from the structure pool.

Application of COMPASS to Solution NMR Data

Though the COMPASS framework was developed to address the problems of spectral overlap and low sensitivity in NMR experiments, it does not rely on any special feature of SSNMR experiments. To test the performance of COMPASS on solution NMR data, a ^{1}H-^{15}N HSQC and a ^{13}C- ^{13}C-^{1}H TOCSY spectra were collected for a uniformly ^{13}C,^{15}N-labeled ubiquitin solution. The 3D TOCSY spectrum was projected through the ^{1}H dimension to generate a ^{13}C-^{13}C 2D spectrum.

The results for the HSQC comparison (Fig 1.11a) do not show a strong relationship between the COMPASS score and the RMSD. We attribute this result to the relative inaccuracy of chemical shift predictions for ^{15}N and ^{1}H amide resonances, due to the stronger dependence on hydrogen bonding and electrostatics, as well as backbone conformation and nearest neighbor residue type. For example, in contrast to the ^{13}Cα predictions which have an RMSD of 0.38 ppm (relative to known chemical shifts for a set of test proteins) (Han et al. 2011), amide ^{15}N predictions have an RMSD of 1.23 ppm, representing a three-fold larger error over a similar range of chemical

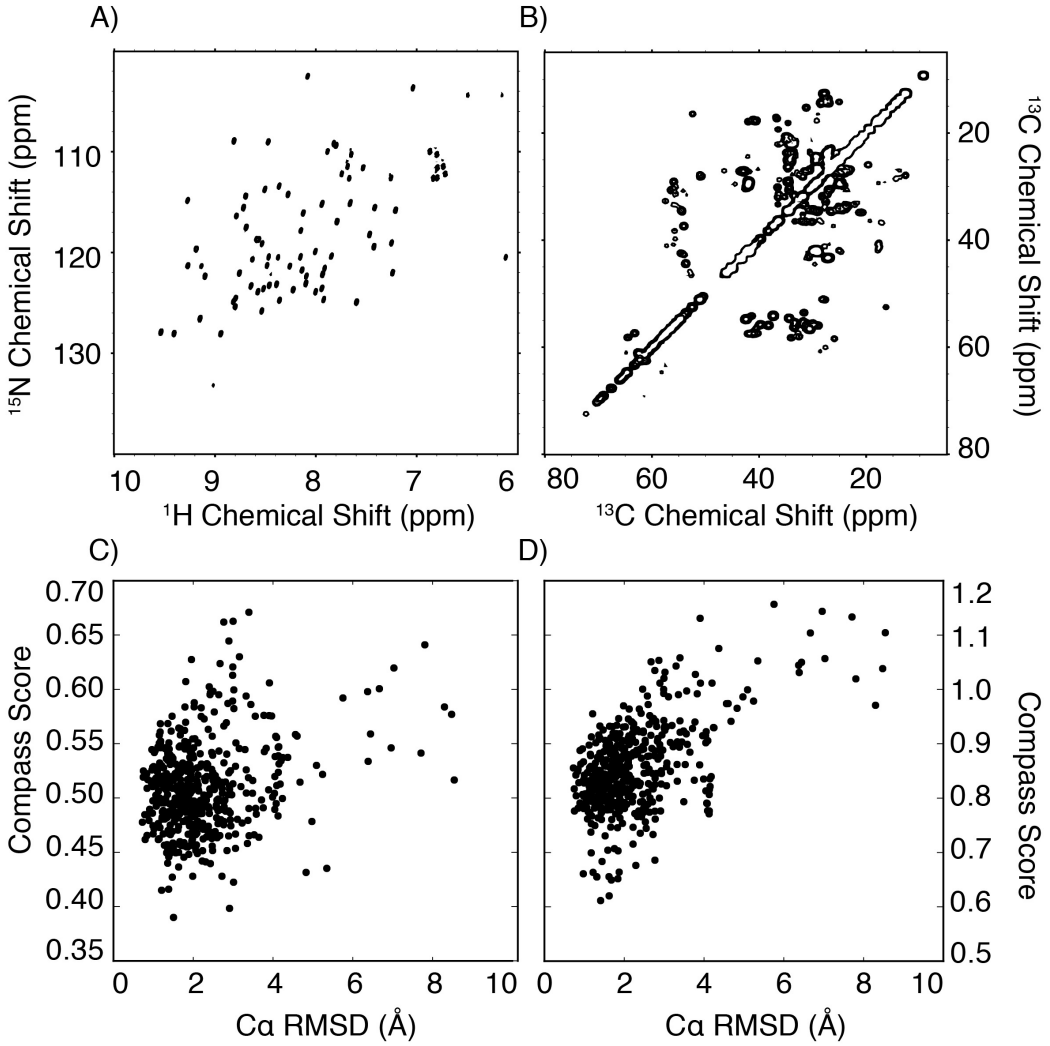

Figure 1.11 - COMPASS applied to solution NMR data of ubiquitin. (A) SOFAST 1H-15N HSQC of ubiqutin (B) F3-projection of 13C-13C-1H TOCSY of Ubiquitin. (C) COMPASS score vs. Cα RMSD for ubiquitin using peaks from HSQC. Difficulty in predicting amide proton and nitrogen shifts makes it unsuited for use with the COMPASS algorithm. (D) COMPASS score vs. Cα RMSD for ubiquitin using peaks from TOCSY spectrum projection. Just as in SSNMR data, the COMPASS score based on 13C-13C correlations has a strong relationship with Cα RMSD allowing its use in the determination of experimentally consistent data.

shifts (~30 ppm overall, or ~6 to 10 ppm for a given residue type). Moreover, the amide ^{1}H shifts have an RMSD of 0.24 ppm over a range of ~3 ppm. Thus the relative error in predicting a ^{1}H-^{15}N correlation spectrum is significantly higher than for ^{13}C-^{13}C spectra, leading in the case of ^{1}H-^{15}N to an inability to conclusively identify the best structure among a set, even for the relatively simple case of ubiquitin.

In contrast, the COMPASS scores for the projected ^{13}C-^{13}C-^{1}H TOCSY spectrum demonstrate a clear correlation and sharp convergence at a low RMSD value (Fig 1.11b), similar to the results observed for the solid-state NMR ^{13}C-^{13}C spectra above, confirming that the strength of this method comes from its use of ^{13}C chemical shifts.

Conclusions

We have presented a new method for objective comparison of a modeled protein structure directly to experimental NMR data. COMPASS greatly reduces the time and effort required to validate a structure with experimental data by circumventing the lengthy process of chemical shift assignment and the collection of large data sets to obtain distance and orientation information required for *de novo* structure determination. The method is robust with respect to data collection and peak picking protocols and has good tolerance for noise and artifacts. Here we have demonstrated successful calculations for 15 proteins, four with experimental SSNMR data, one with experimental solution NMR data, and 10 reconstructed spectra from the BioMagResBank chemical shift database.

The COMPASS algorithm exploits the fact that the ^{13}C chemical shift is an exquisitely sensitive reporter on conformation, including not only backbone conformation as evidenced in the secondary chemical shifts (Spera & Bax 1991), but also the conformation of side chains and packing in the protein core, which give rise to ring current and van der Waals packing effects. COMPASS leverages developments in chemical shift prediction methodology that take these effects into account. Strategies based on empirical models, homology methods, quantum mechanical calculations, and machine learning have progressively improved the accuracy, Here we used SHIFTX2, (Han et al. 2011) which uses a hybrid approach combining a sequence homology module with an ensemble machine learning method to attain good accuracy for both

backbone and side chain atoms. SHIFTX2 attains prediction accuracy of better than 0.6 ppm for α, β, and carbonyl carbons and better than 1.0 ppm accuracy for most side chain carbons. This level of prediction accuracy enables us to use the inherent sensitivity of ^{13}C chemical shifts to discern structural information from NMR data at a much earlier stage of analysis, and to quantitatively judge consistency of raw spectra with structural models. The rapid discrimination of valid protein folds by COMPASS may enable rational prioritization of subsequent data collection for structure refinement and acceleration of data analysis. For example, the experimentally consistent folds identified by COMPASS may be used to perform assignments of ambiguous correlations in spectra with long mixing times, reporting on long-range correlations.

As NMR is applied to systems of increasing complexity, manual data analysis becomes unfeasible. We envision potential future improvements including the application of COMPASS to 3D spectra, the use of the COMPASS score directly in model refinement, and structure determination, as well as continued improvements in the chemical shift prediction accuracy. In the current implementation only ^{13}C chemical shifts are used but to accommodate the inclusion of higher dimensionality data, weighted aggregate scoring functions could be devised to account for differing chemical shift prediction accuracy of different nuclei.

While the combination of MODELLER and SHIFTX works well for the primarily monomeric, globular proteins presented here, the COMPASS algorithm could easily be extended to more specialized areas by using integrative structure prediction approaches for multimeric assemblies (Sali et al. 2015) and utilizing MD averaged chemical shift predictions for dynamic loops (Robustelli et al. 2012). Additionally, our assignment-free approach can be used to replace many chemical shift similarity based potentials for structure refinement and possibly in methods utilizing chemical shifts to develop models of structural ensembles (Kannan et al. 2014).

The continual progression in the quality of model prediction methods and chemical shift prediction algorithms will benefit COMPASS due to its modular approach. By leveraging these increasingly accurate predictions combined with the simple automated analysis of COMPASS previously inaccessible systems will become feasible. These advances may be particularly significant to address categories of proteins, such as membrane proteins and fibrils, which have historically been very challenging.

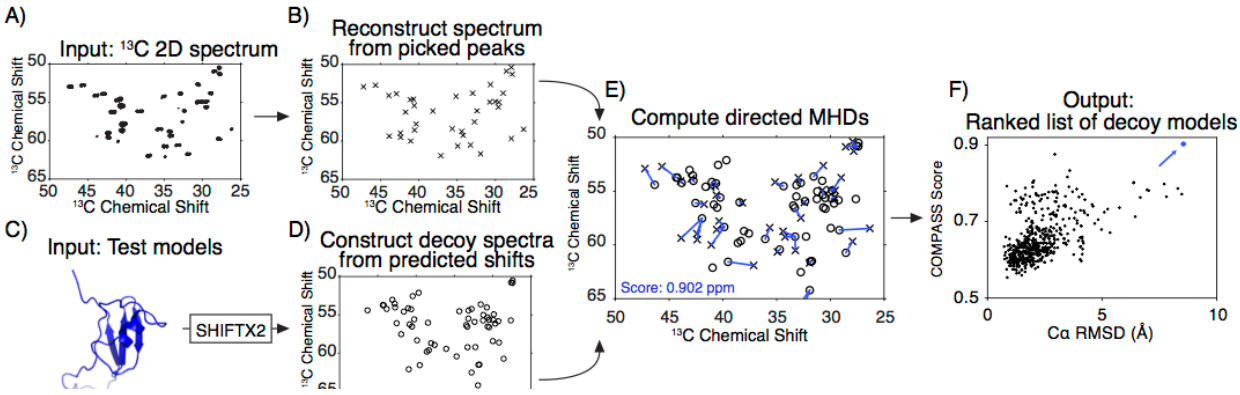

Figure 1.12 - Flow chart of the COMPASS algorithm. (A) The algorithm takes as input a 13C-13C correlation spectrum. A selected region for a spectrum of ubiquitin is shown. (B) The peaks are enumerated and stored as a list of unassigned chemical shift pairs. (C) A collection of test models is produced. The model shown was generated by MODELLER and has a C✔ RMSD of 8.5 Å with respect to the reference structure, 1UBQ. (D) The chemical shifts for each model are predicted by SHIFTX2 and a list of peaks that would occur in a 13C-13C correlation spectrum is generated. (E) The experimental and model peak lists are compared using the COMPASS score. Blue lines indicate the minimum distances described in the text. (F) In this example the COMPASS score from the experimental peak list to the model is 0.902 ppm (point indicated with blue arrow), a relatively high value. The models are then ranked in the order of their computed COMPASS score.

Experimental Procedures

The COMPASS framework can be applied to any combination of model generation method and chemical shift prediction algorithm. In this study, models were prepared using the MODELLER protein structure-modeling program, using a standard protocol, (Eswar et al. 2002) and subsequently relaxed using the *ab initio* relaxation function in the Rosetta software package to ensure low energy side chain conformations. (Simons et al., 1997) SHIFTX2 was used to predict chemical shifts due to its speed and its applicability to both backbone and side chain carbons.

To simulate the 2D spectra, a Python program enumerates all adjacent ^{13}C pairs, assembles the corresponding predicted chemical shifts into pairs and records them in a list (Fig. 1.9). The simulated peak list for each model is then compared to the experimental peak list using the COMPASS score, which is based on the modified Hausdorff distance. Hausdorff distances are a popular family of metrics in computational image analysis and have found applications both in structure comparison and NOESY peak matching. (Zeng et al. 2008, Kozin & Svergun 2001)

The COMPASS score is defined by equations 1 and 2.

$$d(\mathbf{a}, \mathcal{B}) = \min_{\mathbf{b} \in \mathcal{B}} \|\mathbf{a} - \mathbf{b}\|$$

$$d_{COMPASS}(\mathcal{A}, \mathcal{B}) = \frac{1}{\mathcal{N}_{\mathcal{A}}} \sum_{\mathbf{a} \in \mathcal{A}} d(\mathbf{a}, \mathcal{B})$$

Equation 1 defines the distance between a point *a* and a point set *B* as the distance from point *a* to the closest point in set *B*. The COMPASS score is then defined in equation 2 as the average of these minimum distances for every point in set *A*. This definition makes the COMPASS score directional, meaning that switching sets *A* and *B* gives different results. While this diverges from typical Hausdorff distances, it emphasizes the importance of the points in set A (chosen as the experimental peak set) over the points in set B (the predicted peaks). This way, every experimental peak is used in the calculation of the score but if the peak sets are very different, many of the predicted peaks (set B) may be ignored; for example, some regions of a protein may yield lower signal intensities experimentally

The COMPASS score for each model is computed by matching each experimental peak with the nearest predicted peak in the model peak list, and calculating the average minimum distance for these pairings (Fig. 1.10). The COMPASS score is therefore smaller for models that

predict peak patterns similar to the experimental spectrum. In the limit of identical peak patterns, it would be identically zero. By weighting each experimental peak equally, the COMPASS score naturally addresses overlap and missing peaks in experimental spectra. If a peak is missing from the experimental spectrum, nearby peaks in the predicted spectrum are not matched and thus do not contribute to the overall score. Similarly, noise signals are deemphasized by the averaging procedure. Significant outliers that have no near matches in any model peak list contribute a similar magnitude to the scores of all models, manifesting as a nearly constant offset of all resulting scores.

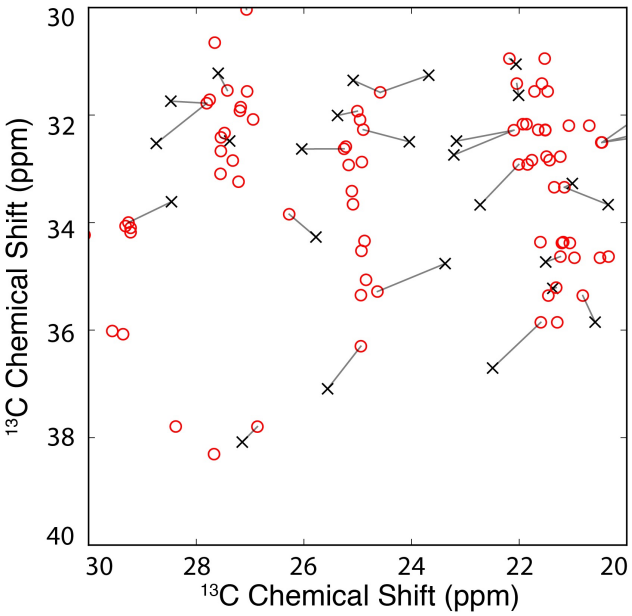

Figure 1.13 - COMPASS score calculation. The COMPASS score is calculated by matching every experimental peak (black x) to the closest test peak (red circle) and calculating the average of the distances between them (gray line). A selected region from a comparison between a ubiquitin COSY spectrum and a poorly matching model is shown.

Sample Preparation

The expression, purification, and crystallization of isotopically labeled recombinant Ubiquitin was previously reported. (Igumenova et al. 2004) The beta 1 immunoglobulin binding domain of protein G (GB1) was expressed and purified as previously reported. (Franks et al. 2005) DsbA was expressed and purified according to the method of Sperling, et al 2010. Soluble Tissue Factor was expressed and purified as described in Boettcher, et al., 2010 and crystallized by precipitation in 1.6 M ammonium sulfate with 200 mM NaCl and 100 mM HEPES buffer pH 7.5

at 4°C as previously reported. (Boys et al. 1993) Samples were packed into 3.2 mm thin wall NMR rotors.

NMR spectroscopy

The ^{13}C-^{13}C 2D CTUC-COSY spectrum of GB1 has been previously reported. (Franks et al. 2005) The CTUC-COSY spectrum of ubiquitin was collected on a 750 MHz Varian VNMRS spectrometer (^{1}H frequency) with a HCN Balun magic angle spinning (MAS) probe. The MAS rate was 16.666 kHz and the variable air temperature was set to -10 °C. SPINAL decoupling (85 kHz) was employed during acquisition. The refocusing delay was 4.2 ms. The spectrum was processed with 20 Hz net line broadening in each dimension.

The CTUC-COSY spectrum of DsbA was collected on a 500 MHz Infinity Plus spectrometer (^{1}H frequency) spinning at 22.222 kHz at VT set point -10 ºC. 85 kHz of 1H SPINAL decoupling was employed during acquisition. 30 Hz net line broadening was applied in each dimension.

The ^{13}C-^{13}C 2D SPC5 spectrum of TF was collected on a 750 MHz Varian VNMRS spectrometer (1H frequency) with a HCN BioMAS probe. The MAS rate was 12.500 kHz and the variable air temperature was set to 10º C. The SPINAL ^{1}H decoupling was employed at 80 kHz during the acquisition. The spectrum was processed with 20 Hz net line broadening in each dimension.

Acknowledgements

This work was performed with Qing Ye, Anna E. Nesbitt, Ming Tang, Marcus D. Tuttle, Eric D. Watt, Kristin M. Nuzzio, Lindsay J. Sperling, Gemma Comellas, Joseph R. Peterson, James H. Morrissey, and Chad M. Rienstra

This work was supported by the National Institutes of Health R01-GM073770 (to C.M.R.) and R01-HL103999 (to J.H.M. and C.M.R.); and NIH S10RR025037 (to C.M.R). J.M.C. and K.M.N. were recipients of National Science Foundation Graduate Research Fellowships. A.E.N. was a recipient of a NIH Ruth L. Kirschstein National Research Service Award (F32 GM095344), and E.D.W. was an American Heart Association Postdoctoral Fellow. The authors thank Dr. Ying Li for expressing and purifying uniformly ^{13}C,^{15}N-labeled ubiquitin and Dr. Deborah A. Berthold for preparing the uniformly ^{13}C,^{15}N-labeled WT-DsbA sample.

References

- Barbet-Massin, E., Pell, A.J., Retel, J.S., Andreas, L.B., Jaudzems, K., Franks, W.T., Nieuwkoop, A.J., Hiller, M., Higman, V., Guerry, P., et al. (2014). Rapid Proton-Detected NMR Assignment for Proteins with Fast Magic Angle Spinning. J. Am. Chem. Soc. *136*, 12489-12497.
- Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. *112*, 535-542.
- Boettcher, J.M., Clay, M.C., LaHood, B.J., Morrissey, J.H., Rienstra, C.M. (2010). Biomol. NMR Assign. *4*, 183-185.
- Boys, C.W.G., Miller, A., Harlos, K., Martin, D.M.A., Tuddenham, E.G.D., O'Brien, D.P. (1993). J Mol Biol *234*, 1263-1265.
- Cavalli, A., Salvatella, X., Dobson, C.M., Vendruscolo, M. (2007). Protein structure determination from NMR chemical shifts. Proc. Natl. Acad. Sci. USA 104, 9615-9620.

- Chen, L., Olsen, R.A., Elliott, D.W., Boettcher, J.M., Zhou, D.H., Rienstra, C.M., Mueller, L.J. (2006). High Resolution (13)C-Detected Solid-State NMR Spectroscopy of a Deuterated Protein. J. Am. Chem. Soc. *128*, 9992-9993.
- Comellas, G., Rienstra, C.M. (2013). Protein Structure Determination by Magic-Angle Spinning Solid-State NMR, and Insights into the Formation, Structure, and Stability of Amyloid Fibrils. Annu. Rev. Biophys. *42*, 515-536.
- Dubuisson, M.P., Jain, A.K. (1994). A Modified Hausdorff Distance for Object Matching. Proc. 12th Intl. Conf. Pat. Recog., 566-568.
- Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M., Pieper, U., Sali, A. (2002).Protein Structure Modeling with MODELLER. In Curr. Prot. in Bioinform. *15*, (John Wiley & Sons, Inc., New York, Unit 5.6), pp 1-30.
- Franks, W.T., Zhou, D.H., Wylie, B.J., Money, B.G., Graesser, D.T., Frericks, H.L., Sahota, G., Rienstra, C.M. (2005). J Am Chem Soc. *127*, 12291-12305.
- Goddard, T.D., Kneller, D.G., SPARKY 3, University of California, San Francisco.
- Guerry, P., Herrmann T. (2011). Advances in automated NMR protein structure determination. Q. Rev. Biophys. *44*, 257-309.
- Güntert, P. (2009). Automated structure determination from NMR spectra. Eur. Biophys. J. *38*, 129–143.
- Han, B., Liu, Y., Ginzinger, S.W., Wishart, D.S. (2011). SHIFTX2: significantly improved protein chemical shift prediction. J. Biomol. NMR *50*, 43-57.
- Hohwy, M., Rienstra, C.M., Jaroniec, C.P., Griffin, R.G. (1999). Fivefold symmetric homonuclear dipolar recoupling in rotating solids: Application to double quantum spectroscopy. J. Chem. Phys. *110*, 7983.

- Hyberts, S.G., Takeuchi, K., Wagner, G. (2010). Poisson-gap sampling and forward maximum entropy reconstruction for enhancing the resolution and sensitivity of protein NMR data. J. Am. Chem. Soc. *132*, 2145-2147.
- Igumenova, T.I., McDermott, A.E., Zilm, K.W., Martin R.W., Paulson, E.K., Wand, A.J. (2004). J. Am. Chem. Soc. *126*, 6720-6727.
- Kannan, A., Camilloni, C., Sahakyan, A.B., Cavalli, A., Vendruscolo, M. (2014). A Conformational Ensemble Derived Using NMR Methyl Chemical Shifts Reveals a Mechanical Clamping Transition That Gates the Binding of the HU Protein to DNA. J. Am. Chem. Soc., *136* (6), 2204–2207.
- Knight, M.J., Webber, A.L., Pell, A.J., Guerry, P., Barbet-Massin, E., Bertini, I., Felli, I.C., Gonnelli, L., Pierattelli, R., Emsley, L., Lesage, A., Herrmann, T., Pintacuda, G. (2011). Fast Resonance Assignment and Fold Determination of Human Superoxide Dismutase by High-Resolution Proton-Detected Solid-State MAS NMR Spectroscopy. Angew. Chem., Int. Ed. *50*, 11697–11701.
- Kozin, M.B., Svergun, D.I. (2001). Automated matching of high- and low-resolution structural models. J. Appl. Crystallogr. *34*, 33-41.
- Linser, R., Bardiaux, B., Andreas, L.B., Hyberts, S.G., Morris, V.K., Pintacuda, G., Sunde, M., Kwan, A.H., Wagner, G. (2014). Solid-State NMR Structure Determination from Diagonal-Compensated Proton-Proton Restraints. J. Am. Chem. Soc. *136*, 11002-11010.
- Lu, J.X., Qiang, W., Yau W.M., Schwieters, C.D., Meredith, S.C., Tycko, R. (2013). Molecular structure of β-amyloid fibrils in Alzheimer's disease brain tissue. Cell *154*, 1257–1268.
- Maly, T., Debelouchina, G.T., Bajaj, V.S., Hu, K.-N., Joo, C.-G., Mak-Jurkauskas, M.L., Sirigiri, J.R., van der Wel, P.C.A., Herzfeld, J., Temkin, R.J., Griffin, R.G. (2008). Dynamic nuclear polarization at high magnetic fields. J. Chem. Phys. *128*, 052211-1-19.

- Moseley, H.N.B., Sperling, L.J., Rienstra, C.M. (2010). Automated Protein Resonance Assignments of Magic Angle Spinning Solid-State NMR Spectra of B1 Immunoglobulin Binding Domain of Protein G (GB1). J. Biomol. NMR *48*, 123–128.
- Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (CASP) - round X. Proteins *82* (S2), 1–6.
- Paramasivam, S., Suiter, C.L., Hou, G., Sun, S., Palmer, M., Hoch, J.C., Rovnyak, D., Polenova, T. (2012). Enhanced Sensitivity by Nonuniform Sampling Enables Multidimensional MAS NMR Spectroscopy of Protein Assemblies. J. Phys. Chem. B *116*, 7416–7427.
- Park, S.H., Das, B.B., Casagrande, F., Tian, Y., Nothnagel, H.J., Chu, M., Kiefer, H., Maier, K., De Angelis, A.A., Marassi, F.M., Opella, S.J. (2012). Structure of the chemokine receptor CXCR1 in phospholipid bilayers. Nature *491*, 779–783.
- Renault, M., Pawsey, S., Bos, M.P., Koers, E.J., Nand, D., Tommassen-van Boxtel, R., Rosay, M., Tommassen, J., Maas, W.E., Baldus, M. (2012). Solid-State NMR Spectroscopy on Cellular Preparations Enhanced by Dynamic Nuclear Polarization. Angew. Chem., Int. Ed. *51*, 2998-3001.
- Robustelli, P., Kohlhoff, K., Cavalli, A., Vendruscolo, M. (2010). Using NMR chemical shifts as structural restraints in molecular dynamics simulations of proteins. Structure. *18*, 923-933.
- Robustelli, P., Stafford, K.A., Palmer, A.G. 3rd. (2012). Interpreting protein structural dynamics from NMR chemical shifts. J. Am. Chem. Soc. 134, 6365-6374.

- Sali, A., Berman, H.M., Schwede, T., Trewhella, J., Kleywegt, G., Burley, S.K., Markley, J., Nakamura, H., Adams, P., Bonvin, A.M.J.J., et al., (2015). Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop, Structure. in press.
- Schmidt, E., Gath, J., Habenstein, B., Ravotti, F., Székely, K., Huber, M., Buchner, L., Böckmann, A., Meier, B.H., Güntert, P. (2013). Automated solid-state NMR resonance assignment of protein microcrystals and amyloids. J. Biomol. NMR *56*, 243–254.
- Shahid, S.A., Bardiaux, B., Franks, W.T., Krabben, L., Habeck, M., van Rossum, B.J., Linke, D. (2012). Membrane-protein structure determination by solid-state NMR spectroscopy of microcrystals. Nat. Methods. *9*, 1212–1217.
- Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J.M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K, Lemak, A., et al. (2008). Consistent blind protein structure generation from NMR chemical shift data. Proc. Natl. Acad. Sci. USA *105*, 4685-4690.
- Simons, K.T., Kooperberg, C., Huang, E., Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J. Mol. Biol. *268*, 209-225.
- Spera, S., Bax, A. (1991). Measurement of NH-CaH coupling constants in staphylococcal nuclease by two-dimensional NMR and comparison with X-ray crystallographic results. J. Am. Chem. Soc. *113*, 5490-5492.
- Sperling, L.J., Berthold, D.A., Sasser, T.L., Jeisy-Scott, V., Rienstra, C.M. (2010). J Mol Biol *399*, 268-282.
- Sun, S., Yan, S., Guo, C., Li, M., Hoch, J.C., Williams, J.C., Polenova, T. (2012). A Timesaving Strategy for MAS NMR Spectroscopy by Combining Non-Uniform Sampling and Paramagnetic Relaxation Assisted Condensed Data Collection. J. Phys. Chem. B *116*, 13585–13596.

- Ulrich, E.L., Akutsu, H., Doreleijers, J.F., Harano, Y., Ioannidis, Y.E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., et al. (2008). BioMagResBank. Nucleic Acids Res. *36*, 402-408.
- Wang, S., Munro, R.A., Shi, L., Kawamura, I., Okitsu, T., Wada, A., Kim, S.-Y., Jung K.-H., Brown, L.S., Ladizhansky, V. (2013). Solid-state NMR spectroscopy structure determination of a lipid-embedded heptahelical membrane protein. Nat. Methods. *10*, 1007–1012.
- Wang, T., Park, Y.B., Caporini, M.A., Rosay, M., Zhong, L., Cosgrove, D.J., Hong, M. (2013). Sensitivity-enhanced solid-state NMR detection of expansin's target in plant cell walls. Proc. Natl. Acad. Sci. USA *110*, 16444-16449.
- Wasmer, C., Lange, A., Van Melckebeke, H., Siemer, A.B., Riek, R., Meier, B.H. (2008). Amyloid fibrils of the HET-s(218–289) prion form a beta-solenoid with a triangular hydrophobic core. Science *319*, 1523–1526.
- Zeng, J., Tripathy, C., Zhou, P., Donald, B.R. (2008). A Hausdorff based NOE assignment algorithm using protein backbone determined from residual dipolar couplings and rotamer patterns. Comput. Sys. Bioinform. Conf., 169–181.
- Zhou, D.H., Nieuwkoop, A.J., Berthold, D.A., Comellas, G., Sperling, L.J., Tang, M., Shah, G.J., Brea, E.J., Lemkau, L.R., Rienstra, C.M. (2012). Solid-state NMR analysis of membrane proteins and protein aggregates by proton detected spectroscopy. J. Biomol. NMR *54*, 291–305.

CHAPTER 2: An Analytical Expression for NMR Observable Uncertainty

Introduction

It is common practice to analyze NMR data in the frequency domain, most commonly by apodizing and Fourier transforming the data. The choice of apodization parameters is somewhat of an art form and an experienced spectroscopist will choose parameters to accentuate the aspects of the data that are most important for the question at hand. However, when deriving measurements from the frequency domain data, the error analysis is often very simplistic.

Typically, errors in amplitudes are estimated as equal to the root mean square noise, errors in the frequencies are estimated as equal to the linewidths, and errors in the linewidths are rarely considered at all. These errors do not incorporate the effects of apodization, the sampling schedule, or the model used for peak fitting. Below, I give a theoretical analysis of the uncertainty in parameters derived from frequency domain data as determined by the apodization parameters and the model function used in peak fitting. Then, I expand the analysis to include non-Fourier methods of reconstruction, which is especially important at a time when Fourier methods are being replaced by more complex spectral reconstruction techniques necessitated by the increasing popularity of non-uniform sampling of time-domain data.

Modelling Noisy Signals

A signal is a set of measurements with some underlying structure relating them. All real signals are stochastic, whether the values themselves are the result of a stochastic process, like a speck of pollen undergoing Brownian motion in water, or are corrupted by a stochastic process, like an electrical signal containing thermal noise. For many signals, the uncertainty over any single measurement is best described by a normal distribution. When a large number of these

measurements are considered together they can be described by a Gaussian process, a possibly infinite-dimensional generalization of the normal distribution. Formally, a Gaussian process is a collection of random variables, any subset of which has a joint Gaussian distribution. Often Gaussian processes are described using a mean function mu(x) and a covariance function $k(x_i,x_j)$. A particularly fruitful interpretation [Rasmussen, 2006] of Gaussian processes is as a distribution over functions. In this interpretation a single draw from a Gaussian process is a function $f(x) \sim GP(\mu(x), k(x_i,x_j))$. The mean function, $\mu(x)$, describes the average value of all functions f at each value of x and the covariance function, $k(x_i,x_j)$, describes the correlation between any two values of the functions f at positions x_i and x_j.

As an example, let's consider a Gaussian process $GP(0,k(x_i,x_j))$ where the covariance function (or covariance kernel) is the squared exponential kernel.

$$k(x_i, x_j) = \sigma^2 e^{-\frac{(x_i - x_j)^2}{\tau^2}}$$

Upon inspection, it is clear that the covariance of the function where xi=xj is equal to the variance and that the covariance of the function decreases as the distance between the values of xi and xj increases. Intuitively this indicates that nearby function values will be very similar to each other. Observing a collection of draws from this Gaussian process in Figure 2.1, we see that this intuition is correct. The functions are all smoothly varying and vary around the mean value of zero. We can plot multiples the standard deviation, $\sigma(x)=sqrt(k(x,x))$ as a shaded region around the mean function to indicate the probability of any individual function value existing at that distance from the mean. As expected, the functions drawn from the Gaussian process rarely foray more than three standard deviations from the mean. (Fig. 2.1)

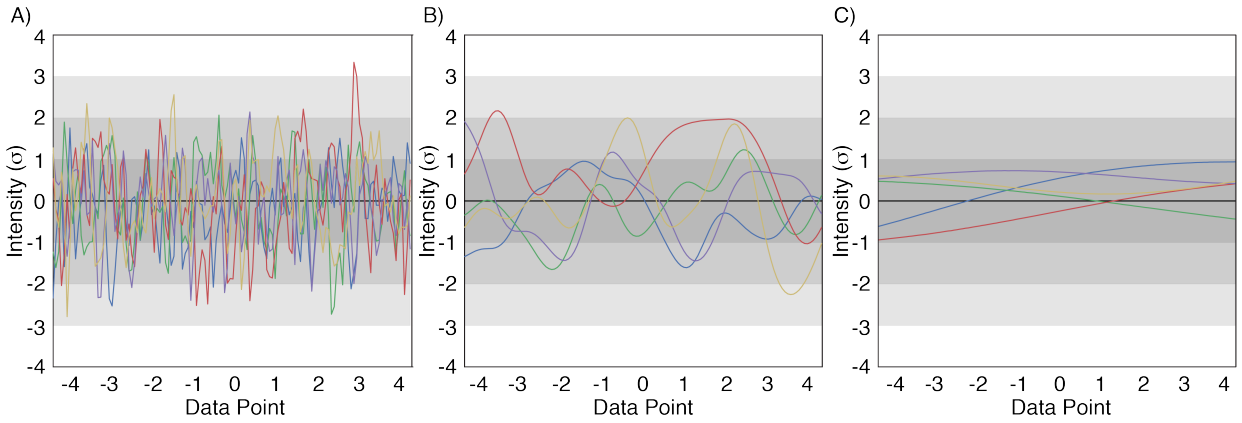

Figure 2.1 - Draws from a Gaussian process with squared-exponential covariance with varying length-scales. A) length scale 0.1. B) length scale 1.0. C) length scale 10.0

The properties of the drawn functions vary drastically with the form of the covariance function. For instance, varying the length scale parameter, tau, in the squared exponential function, generates distributions over signals with very different characters. (Fig. 2.1). When the length scale is very long, the functions are effectively constant; when the length scale is extremely short (approximating a delta function), we recover uncorrelated, white Gaussian noise.

Gaussian processes are especially useful because their mathematical form makes them extremely easy to work with. For instance, for Gaussian processes (or more generally, any continuous process), the Wiener-Khinchin theorem [Wiener, 1964] states that

$$k(\tau) = \int S(f)\, e^{2\pi i \tau f}\, df$$

or conversely,

$$S(f) = \int k(\tau) e^{-2\pi i \tau f}\, d\tau$$

In less precise terms, the Wiener-Khinchin theorem states that all valid correlation functions are Fourier transforms of continuous, square-integrable functions and that the correlation function fully describes the spectral density of draws from the process.

Using the two forms of the Wiener-Khinchin theorem, we will be able to analyze the effect of transformations applied to a Gaussian process in the time domain and determine their effects on the frequency domain.

A Stochastic Model of NMR Signals

A time-domain NMR signal may be modelled as a sum of decaying sinusoids, corrupted with uncorrelated, white, Gaussian noise. If we rephrase this model in the language of Gaussian

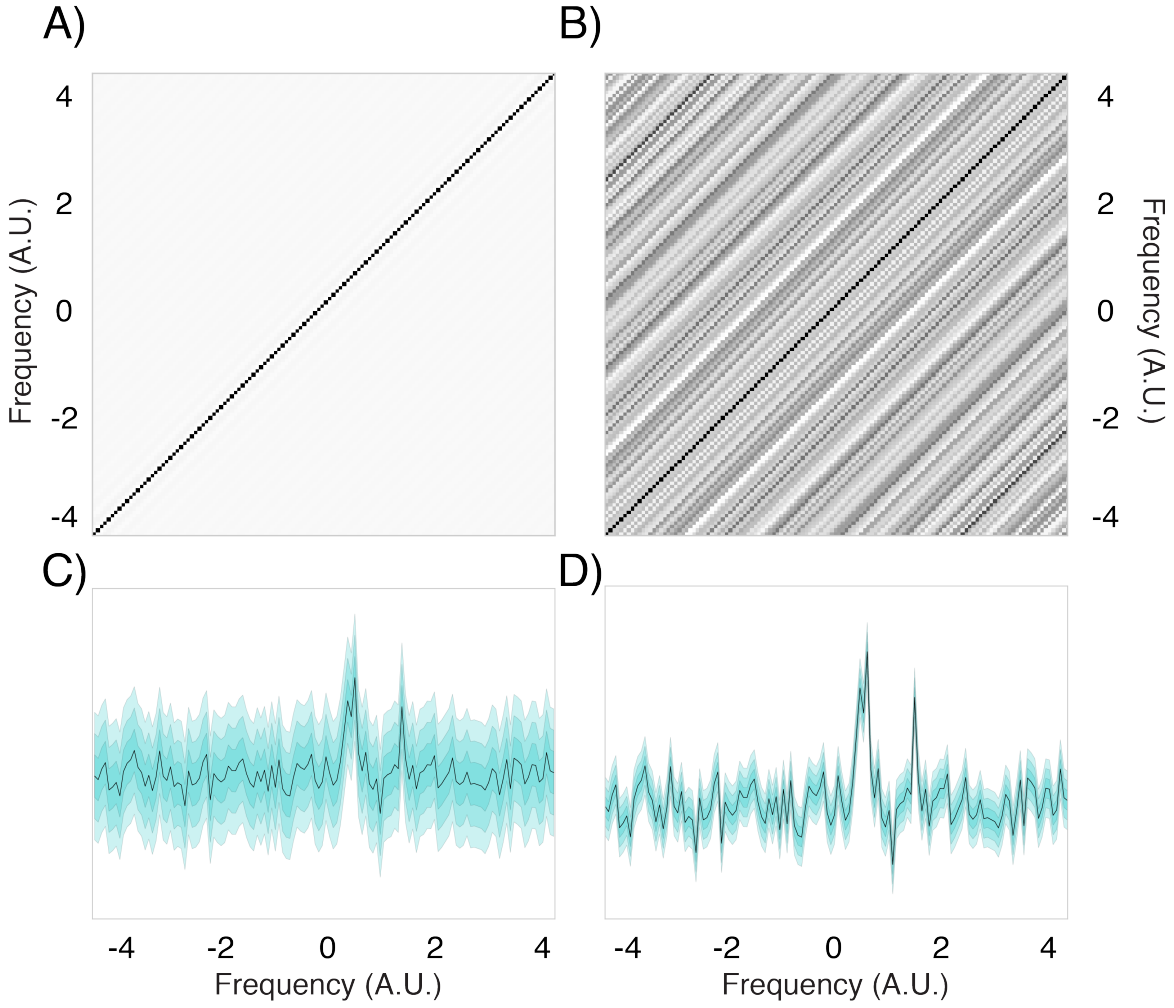

Figure 2.2 - Frequency domain covariance structure for uniformly and non-uniformly sampled data. a-b) frequency-frequency covariance matrices. c-d) spectra with 1,2, and 3 sigma regions shaded. 128 points sample in uniform spectrum, 11 in non-uniform.

processes, a time-domain NMR signal is a Gaussian process with a mean function consisting of a sum of decaying sinusoids and a covariance function $k(t_i,t_j) = \delta(t_i,t_j)*\sigma$. Now we would like to analyze the effect of apodization on the noise present in the data.

Apodization is, generally, the multiplication of a signal by another function prior to Fourier transformation. It is often motivated by emphasizing certain aspects of the signal that are most interesting to the researcher. For instance, weighting a signal with a decaying sinusoid with a decay constant equal to the natural decay rate of the coherent portion of the signal results in maximized signal-to-noise ratio at the expense of increased linewidths. We can determine the effect this exponential weighting has on the covariance structure of the noise in the frequency domain by application of the Wiener-Khinchin theorem [Wiener, 1964], specifically

$$K_f(\nu) = S(\nu) = \int k(\tau)\, e^{-2\pi i \nu \tau} d\tau$$

Here we have interpreted the spectral density as the covariance function of the Gaussian process represented in the frequency domain and use the substitution $nu=f_i-f_j$. The apodization takes the form of a multiplication of the time-domain covariance function which now takes the form $k(\tau) = \exp(-\tau/t_a)$. Upon integration, we find that the frequency domain covariance function is the familiar Lorentzian function. The effect of this Lorentzian covariance function is that the noise in the frequency domain is smoothed and nearby points are no longer independent.

The application of the continuous form of the Wiener-Khinchin theorem does not provide substantial new insight into the effect of apodization that could not be derived from the traditional convolution interpretation. Nevertheless, the application of the discrete Wiener-Khinchin theorem to non-uniformly sampled data reveals a rarely discussed effect of non-uniform sampling.

The discrete Wiener-Khinchin theorem can be stated as:

$$K_f(f_i, f_j) = \sum_k k(\tau)\, e^{-2\pi i (f_i - f_j) t_k}$$

When applied to non-uniformly sampled data, the resulting frequency-domain covariance function takes on much more complex structure (Fig. 2.2a-b). Counterintuitively, the uncertainty in the reconstruction of the spectrum is actually smaller than in the uniformly sampled case. This is due to the somewhat unrealistic assumption that the method used to reconstruct the spectrum is artifact-free. The uncertainty shown here is due solely to noise in the time-domain measurements and since the reconstructed spectrum has the same signal intensity but contains fewer noisy datapoints the noise is substantially lower. The real impact of non-uniform sampling is in the loss of independence between datapoints in the spectrum at substantially different frequencies. While the covariance function for the uniformly sampled data is concentrated near the diagonal indicating only local correlations between datapoints, the non-uniformly sampled spectrum has large covariance values for most pairs of datapoints ranging across the entire spectrum. This loss of independence has been the cause of a large amount of confusion in the non-Fourier spectral reconstruction literature, because it makes simple calculations of signal-to-noise ratios invalid. In the next section, we proceed to explore the consequences of a complex covariance function on the uncertainty in the determination of observables from NMR data.

A General Formula for Uncertainty in Spectral Observables

The measurement of useful observables from NMR usually involves the identification and fitting of peaks. While there are many heuristics for this approach, one with the strongest theoretical background is non-linear least squares (NLS) regression. From the standard statistical treatment, we know that the covariance among the parameters determined by NLS regression is given by [Taylor, 1997]

$$K_{par} = J^* K_{meas} J$$

Here J represents the Jacobian matrix of the model with respect to variations in the model function and k_meas represents the covariance matrix in the raw data. If we adopt a model of the peak shape in our data, we can calculate the Jacobian for a given set of frequency values and fit parameters to determine error estimates for the amplitude, frequency, and linewidth of peaks in our spectrum.

An interesting connection to the Wiener-Khinchin theorem above can be seen if we recognize that when using non-Fourier reconstruction of spectra from sparse, non-uniformly sampled time-domain data, the spectrum is no longer an equivalent representation of the raw time-domain data but more of a set of parameters of a complex nonlinear model. If we consider the amplitudes in a reconstructed spectrum to be parameters in a model fit to experimental data, we can apply the above equation relating the measurement covariance in the time domain to the covariance in the frequency domain. If we assume that the method used to reconstruct the spectrum from the NUS time-domain data is optimal and provides an unbiased estimate of the true spectrum, then we can adopt the inverse discrete Fourier transform as our nonlinear model. In calculation of the Jacobian matrix, we recognize that the elements are simply complex exponentials for each pair of frequency increment and time measurement. This means that the Jacobian matrix is effectively a DFT. Thus, uncertainty analysis by way of Gaussian process analysis and NLS regression give equivalent results.

By combining the error analysis in the spectral reconstruction and the fitting of peak line shapes, we can develop a hierarchical model relating apodization and non-uniform sampling to uncertainty in peak position, height, and width. Combining

$$K_{peak} = \left[J_{peak}^* K_{spec} J_{peak}\right] \text{and } K_{spec} = \left[J_{spec}^* K_{meas} J_{spec}\right]$$

We are left with:

$$K_{peak} = \left[J^*_{peak} J^*_{spec} K_{meas} J_{spec} J_{peak}\right]$$

where

$$K_{meas} = diag\left(f_{apod}\right) \times \sigma$$

This formula is the major result of this chapter and relates all the experimental parameters related to sampling, apodization, and the peak model to the error and covariance among the parameters of the peak model.

Results and Discussion

Preliminary results are given below. We plan computational experiments to test the limits of the uncertainty analysis presented in this chapter. Figure 2.3 demonstrates the effect the random seed used when constructing a sampling schedule can have on the ability to measure a frequency

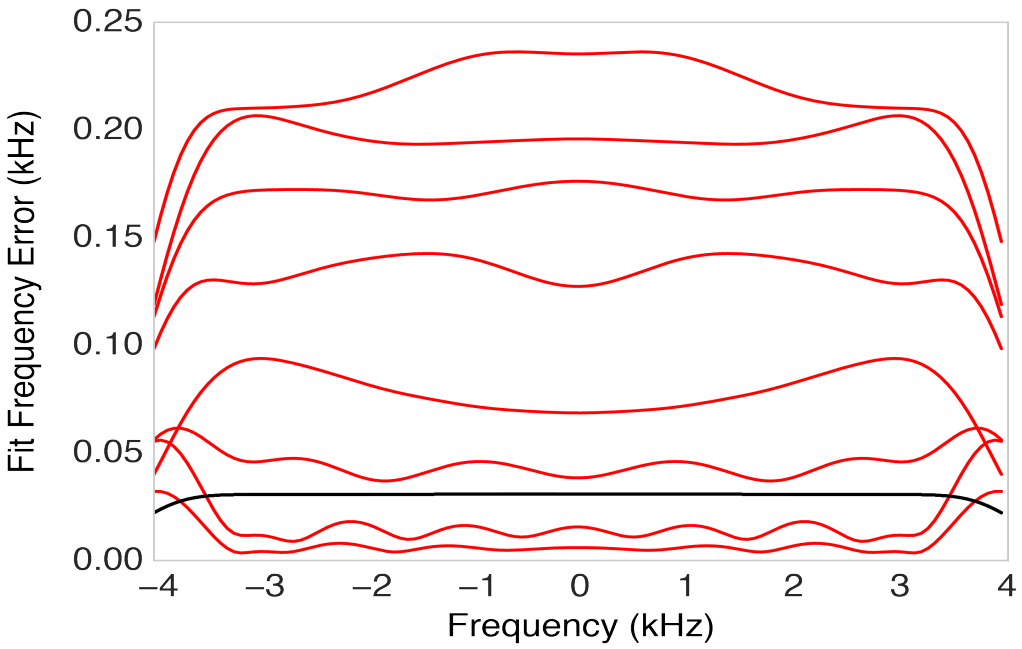

Figure 2.3 - Error in frequency estimate as a function of frequency of peak. Error for a uniformly sampled spectrum is shown in black. Error for non-uniformly sampled spectra sampled at 11 poitns with different sampling schedules are shown in red.

accurately, independent of the method used to reconstruct the spectrum. The standard considerations when choosing a uniform sampling schedule are based on the sensitivity-resolution-time tradeoff introduced by Fourier analysis. However, the tradeoffs become more complex and, until now, more difficult to quantify when moving to non-uniform sampling schedules. As can be seen below, although the spectrum would require much less time to collect (11 vs 128 points) and, presuming the reconstruction method is successful, would have the same signal but lower noise content leading to a higher signal-to-noise ratio, the resulting fit frequencies of peaks in the spectrum are usually of higher uncertainty.

The analysis given here examines the effect of apodization and sampling on noise and the introduction of correlations between points typically considered to be independent in spectral reconstructions. Independent of the content of the spectrum it is possible to determine the effects of these experimental choices on the ability to extract meaningful information from the data. A more complete analysis comparing uniform and non-uniformly sampled data is underway to better understand the limits of the formulas presented above.

References

- Carl Edward Rasmussen and Chris Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006.
- Norbert Wiener. Time Series. M.I.T. Press, Cambridge, Massachusetts. 1964.
- John Robert Taylor. An Introduction to Error Analysis. University Science Books. 1997.

CHAPTER 3: Model-Free Fitting of Dipolar Coupling Trajectories

Introduction

^{1}H-^{13}C dipolar couplings are rich sources of structural and dynamical information. The value of the dipolar coupling depends on geometry, specifically the relative orientations of the two spins and the external magnetic field (equation 1) [Schmidt-Rhor, 1994]. However, in solution NMR, isotropic tumbling faster than the typical timescale of dipolar couplings (10-100 kHz) leads to a complete averaging of the interaction causing it to be unobservable to first order in spectra. This averaging has a major benefit in that it leaves spectral line shapes narrow, only depending on the relaxation properties of the spin. In magic angle spinning solid-state NMR (MAS SSNMR) a similar averaging is achieved by mechanical rotation of the sample around an axis oriented at the "magic angle" relative to the magnetic field at rates comparable to the dipolar couplings. In solution NMR it is desirable to recover a fraction of the global orientation dependence and this can be done by causing partial alignment of the protein solution by the introduction of alignment media in the sample such as strained polyacrylamide gel [Bax, 2005]. Partial alignment leads to a slight bias in rotational diffusion causing an incomplete averaging of the dipolar interaction which can be observed as a splitting in spectral lines. The weak RDCs are adequately approximated as two-spin interactions and conveniently allow a simple, direct relationship between the measured coupling and the orientation of the interatomic vector relative to the alignment medium.

$$\hat{H}_D = \frac{\mu_0 \gamma_I \gamma_S h^2}{16\pi^3 r^3} \left[\left(\hat{\vec{I}} \cdot \hat{\vec{S}} \right) - 3 \frac{\left(\hat{\vec{I}} \cdot \vec{r} \right) \left(\hat{\vec{S}} \cdot \vec{r} \right)}{r^2} \right] \quad (1)$$

In MAS SSNMR it is also possible to partially recover the dipolar coupling interaction but since the averaging under MAS is coherent, recoupling is achieved through synchronous radio frequency pulses with phases and timings that exploit the inherent symmetry of the rotating

frame Hamiltonian to interfere with the averaging process [Levitt, 2002]. Symmetry-based sequence recoupling has the advantage over partial alignment methods in that it allows the dipolar couplings to be "turned on" only when needed so that the chemical shifts of the involved spins can be measured independently of the dipolar couplings. It turns out that this is necessary in solid-state NMR because recoupled dipolar couplings are a few orders of magnitude stronger than those in solution RDCs, so that they can be measured before the involved spins relax due to the much shorter transverse relaxation time in MAS SSNMR. The resulting strongly coupled network of spins leads to a loss of the simple two-spin dynamics observed in solution and necessitates modeling of multi-spin dynamics that cannot be described by simple analytical relationships such as

$$\hat{H}_D = \frac{\hbar\gamma_I\gamma_S}{4\pi^2 r^3}\left[1 - 3cos^2\theta\right]\left(3I_zS_z - \vec{I}\cdot\vec{S}\right) \quad (2)$$

Rather, a matrix of couplings must be constructed in order to numerically evaluate the spin dynamics under several couplings.

A major benefit of the multi-spin behavior of dipolar evolution in MAS SSNMR experiments is that the relative orientations of dipolar couplings leads to distinct behavior allowing the measurement of bond angles and dihedral angles with high precision. That measurement is complicated, however, by incoherent dynamic averaging of dipolar couplings due to molecular motion. Considering the secular dipolar coupling of two spins (equation 2), one observes that both distance dependence and the orientational dependence enters the equation as a scaling factors of r^{-3} and $1\text{-}3\cos^2(\theta)$, respectively. Dynamic averaging of the angular factor leads to a scaling of the dipolar coupling by a factor with absolute magnitude anywhere between 0 and 1, a quantity known as an order parameter. Because the order parameter and the distance enter the equation as a product,

they cannot be separated without the inclusion of external information or global fitting of multiple dipolar couplings involving each of the two coupled spins.

The difficulty of interpretation of recoupled dipolar couplings has traditionally led to the adoption of restrictive models that assume geometric restraints on relative coupling orientations or tie the values of parameters on multiple atoms to be equal. Approximations such as these sometimes lead to incorrect interpretations and often lead to low-quality fits. To address these issues, we developed a model-free fitting procedure that treats the trajectory of single quantum amplitude of a single S spin surrounded by a constellation of I spins without placing restrictions on the relative positions of the I spins or placing equivalence constraints on their parameters. We employ an average Liouvillian theory-based treatment following previous approaches [Hohwy 2000; Rienstra, 2002] but with a fitting procedure that is more robust, allows a more detailed description of the spin interactions, and allows us to interpret previously uninterpretable dipolar trajectories.

Theory

Under MAS and the application of a symmetry-based pulse sequence designed to recouple heteronuclear dipolar couplings but decouple homonuclear dipolar couplings for both the I and S spins, the dipolar coupling Hamiltonian (equation 1) is transformed into an average Hamiltonian (equation 3) that is a valid approximation for the spin interactions during integer multiples of the basic element of the recoupling sequence [Schmidt-Rohr, 1994]. Since a given sample in SSNMR consists of many individual molecules all at different orientations, the different orientational populations or “crystallites” must be described separately The recoupled Hamiltonian describes the orientation dependence through the use of Wigner rotation matrices of order two, which

describe the rotation of the spin-part and spatial parts of the dipolar coupling tensor in terms of Euler angles relative to three reference frames: the rotor frame, the crystallite frame, and the bond frame or more commonly "molecular" frame even though each spin cluster in a molecule may have an independent orientation in modelling. The Wigner rotation matrices can be separated into the effects of each angle into complex exponential factors for the alpha and gamma angles and more complex relationship for beta given by the rank-2 Wigner d-matrix with values given in table

$$H_{IS} = \sum_{i}^{n} \sum_{j}^{k} \left(\omega_{I_i,S_j}^{(1)} I_i^{+} + \omega_{I_i,S_j}^{(-1)} I_i^{-} \right) S_{jZ} \tag{3}$$

$$\omega_{I_i,S_j}^{(1)} = \left(\omega_{I_i,S_j}^{(-1)} \right)^{\star} = -\frac{2\pi}{\sqrt{3}}\, a\, b_{IS} \sum_{\substack{m=-2 \\ m\neq 0}}^{2} D_{-1,-m}^{2} \left(\alpha_{RC}, \beta_{RC}, \gamma_{RC} \right) D_{-m,0}^{2} \left(0, \beta_{CB}, \gamma_{CB} \right)$$

$$b_{IS} = \frac{\gamma_I \gamma_S \hbar}{r_{IS}^{3}} \frac{\mu_0}{4\pi}$$

Here we have separated the parameters of the dipolar coupling into three sets, the crystallite-to-principle axis system aligned with the interatomic vector Euler angles, the rotor-to-crystallite Euler angles, and a general scaling factor, a, which contains the dynamic and distance dependence of the coupling. The constant factor b_{IS} represents the magnitude of the dipolar coupling of a perfectly rigid pair of I and S spins at 1 Å distance scaled by the scaling constant of the recoupling sequence. The averaging over the rotor rotation that cannot be recovered is captured in the factor of 1/sqrt(3) and the factor of 2pi puts the dipolar coupling in units of radians/s.

$$\frac{d|\sigma(t)\rangle}{dt} = -\left(i\hat{H} + \Gamma \right) |\sigma(t)\rangle \tag{4}$$

The full time-dependent Schrödinger equation involving this Hamiltonian is not analytically solvable with an arbitrary number of spins, therefore we take the standard approach of average Liouvillian theory (ALT) and approximate the Hamiltonian as a sum of a purely oscillatory component H, and a diagonal relaxation super-operator Gamma (equation 4). Although the dipolar couplings have complex values, their phase has no observable effect and thus their absolute values appear in the ALT matrix; the vertical bars are omitted for visual clarity. This ALT operator acts on the time dependent state vector sigma to describe the observed change in the state vector. Using basic differential equations, it is straightforward to see that the solution to equation 4 is given by equation 5, which describes the evolution of the system as the matrix exponential of the ALT operator acting on the initial state vector. The matrix form of the differential equation is given in equation 6 showing the sparse structure of the ALT matrix.

$$|\sigma(t)\rangle = e^{\left(i\hat{H}+\Gamma\right)t}|\sigma_0\rangle \tag{5}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{vmatrix} S_{jX} \\ 2S_{jY}I_{1\psi} \\ 2S_{jY}I_{2\psi} \\ 2S_{jY}I_{3\psi} \\ -4S_{jX}I_{1\psi}I_{2\psi} \\ -4S_{jX}I_{1\psi}I3_{\psi} \\ -4S_{jX}I_{2\psi}I_{3\psi} \\ -8S_{jY}I_{1\psi}I_{2\psi}I_{3\psi} \end{vmatrix} = \begin{vmatrix} -\Gamma_1 & -\omega_{j1} & -\omega_{j2} & -\omega_{j3} & & & & \\ \omega_{j1} & -\Gamma_2 & & & -\omega_{j2} & -\omega_{j3} & & \\ \omega_{j2} & & -\Gamma_2 & & -\omega_{j1} & & -\omega_{j3} & \\ \omega_{j3} & & & -\Gamma_2 & & \omega_{j1} & -\omega_{j2} & \\ & \omega_{j2} & \omega_{j1} & & -\Gamma_3 & & & \omega_{j3} \\ & \omega_{j3} & & \omega_{j1} & & -\Gamma_3 & & -\omega_{j2} \\ & & \omega_{j3} & \omega_{j2} & & & -\Gamma_3 & -\omega_{j1} \\ & & & & \omega_{j3} & \omega_{j2} & \omega_{j1} & -\Gamma_4 \end{vmatrix} \begin{vmatrix} S_{jX} \\ 2S_{jY}I_{1\psi} \\ 2S_{jY}I_{2\psi} \\ 2S_{jY}I_{3\psi} \\ -4S_{jX}I_{1\psi}I_{2\psi} \\ -4S_{jX}I_{1\psi}I_{3\psi} \\ -4S_{jX}I_{2\psi}I_{3\psi} \\ -8S_{jY}I_{1\psi}I_{2\psi}I_{3\psi} \end{vmatrix} \tag{6}$$

In typical usage, only the single-quantum coherence of the S spin is translated into observable signal and thus the signal is described as the projection of the evolved state vector onto a measurement vetor p = (1 0 0 … 0). Similarly, the spin cluster is typically prepared in the single quantum state and signals originating from other coherence orders are filtered out before detection leading to an initial state described by a vector identical to the measurement vector, σ(0) = [1 0 0 … 0]. As the observed signal is the sum of the contributions of each crystallite, the predicted signal

is described as a sum of the individual signals over all crystallite orientations, which leaves us with a total of three free parameters for describe each dipolar coupling.

To fit these dipolar couplings, we use an iterative procedure of simulation and least-squares comparison. Because of the many symmetry-related solutions, the optimization process is not convex and simple gradient-descent methods become stuck in local minima. Additionally, because the relevant number of coherence orders scales as 2^n where n is the number of I spins, the simulation quickly slows with the spin clusters of increasing size and grid search over the parameters is infeasible for clusters larger than a handful of spins. Therefore we chose to use an optimization procedure known as differential evolution.

Differential evolution is an evolutionary optimization algorithm that searches parameter space by marinating a population of candidate solutions that move around the search space by a combination of random movement and crossover which combines the parameter values of different candidates to produce new solutions. The algorithm proceeds, leading to populations of greater average quality while still searching the full parameter space until a final candidate is chosen. Often, due to the stochastic nature of the algorithm, the resulting parameter set is within the basin of the global optimum but is slightly elevated above the optimum and therefore we "polish" the solution found by differential evolution by a final gradient descent optimization.

Error propagation

Once we have found an optimal solution, to perform error analysis we make the assumption that small perturbations of the dipolar trajectory due to noise in the measurement will not move the optimal solution out of the local minimum and as such we can approximate the local minimum of the error surface as a quadratic function which leads to uncertainty in parameter values that are

linear in experimental noise. The parameter errors are related to the experimental noise by the derivative of the parameter with respect to small perturbations in the signal.

We desire to find the derivative of the parameter values with respect to measured signal values. As the error function is not solvable for the parameter values, we make use of the inverse relationship between dx/dS and dS/dx and differentiating the error function

$$S = \sum_{i=0}^{n} \left(y_i - s\left(t_i|\Gamma_2, a_1, \beta_{CB,1}, \gamma_{CB,1}, \beta_{CB,2}, \gamma_{CB,2}, \cdots\right)\right)^2 \tag{7}$$

With respect to the measured signal values giving us

$$\frac{\partial S}{\partial x} = -2\sum_{i=0}^{n} \frac{\partial s\left(t_i|x\right)}{\partial x}\left(y_i - s\left(t_i|x\right)\right) \tag{8}$$

By application of the product rule we find the error derivative as sum of derivatives of the simulated trajectory with respect to the parameters. Further application of the product rule gives us expansions of the derivatives of the simulated trajectory with respect to each parameter

$$A = -\left(i\hat{H} + \Gamma\right)$$

$$\frac{\partial s\left(t\right)}{\partial \beta_{CB}} = \frac{\partial s\left(t\right)}{\partial A}\frac{\partial A}{\partial \omega_{I_i,S_j}}\frac{\partial \omega_{I_i,S_j}}{\partial \beta_{CB}} \qquad \frac{\partial s\left(t\right)}{\partial \gamma_{CB}} = \frac{\partial s\left(t\right)}{\partial A}\frac{\partial A}{\partial \omega_{I_i,S_j}}\frac{\partial \omega_{I_i,S_j}}{\partial \gamma_{CB}}$$

$$\frac{\partial s\left(t\right)}{\partial r} = \frac{\partial s\left(t\right)}{\partial A}\frac{\partial A}{\partial \omega_{I_i,S_j}}\frac{\partial \omega_{I_i,S_j}}{\partial r} \qquad \frac{\partial s\left(t\right)}{\partial a} = \frac{\partial s\left(t\right)}{\partial A}\frac{\partial A}{\partial \omega_{I_i,S_j}}\frac{\partial \omega_{I_i,S_j}}{\partial a}$$

$$\frac{\partial s\left(t\right)}{\partial \Gamma_2} = \frac{\partial s\left(t\right)}{\partial A}\frac{\partial A}{\partial \Gamma_2} \tag{9}$$

Where A is the negative ALT matrix.

Recognizing that the trajectory is described by an equation of the same form

$$\frac{\partial}{\partial A}x'e^{At}z = \begin{bmatrix}\mathbf{0}_r & \mathbf{I}_r\end{bmatrix} e^{Ct} \begin{bmatrix}\mathbf{I}_r \\ \mathbf{0}_r\end{bmatrix}$$
$$C = \begin{bmatrix}A' & \mathbf{0}_r \\ xz' & A'\end{bmatrix} \tag{10}$$

We can use a result in [Willy, 2008] that describes the derivative of a matrix exponential sandwiched between two projection vectors

$$s(t) = \vec{p}\, e^{-\left(i\hat{H}+\Gamma\right)t}\sigma(0) \tag{11}$$

Allowing us to calculate the derivatives of the ALT matrix with respect to the dipolar couplings

$$\frac{\partial A}{\partial \omega_{I_i,S_j}} = \begin{cases} 1 & \text{if } \sigma_n \setminus \sigma_m = I_i, n > m, \\ -1 & \text{if } \sigma_n \setminus \sigma_m = I_i, n < m, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

And relaxation rate

$$\frac{\partial A}{\partial \Gamma_2} = \begin{cases} -\frac{\Gamma_i}{\Gamma_2} & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

Using the derivative of the absolute value of a complex valued function, the derivatives of the dipolar coupling with respect to the scaling parameter a is given by the simple form

$$\begin{aligned}\frac{\partial \omega_{I_i,S_j}}{\partial a} &= \frac{1}{|\omega^{(1)}_{I_i,S_j}|}\left(\mathrm{Re}\left[\omega^{(1)}_{I_i,S_j}\right]\mathrm{Re}\left[\frac{\partial \omega^{(1)}_{I_i,S_j}}{\partial a}\right] + \mathrm{Im}\left[\omega^{(1)}_{I_i,S_j}\right]\mathrm{Im}\left[\frac{\partial \omega^{(1)}_{I_i,S_j}}{\partial a}\right]\right) \\ &= \frac{1}{|\omega^{(1)}_{I_i,S_j}|}\left(\mathrm{Re}\left[\omega^{(1)}_{I_i,S_j}\right]\mathrm{Re}\left[\omega^{(1)}_{I_i,S_j}\right] + \mathrm{Im}\left[\omega^{(1)}_{I_i,S_j}\right]\mathrm{Im}\left[\omega^{(1)}_{I_i,S_j}\right]\right) \\ &= |\omega^{(1)}_{I_i,S_j}| = \omega_{I_i,S_j}\end{aligned} \tag{14}$$

Whereas the derivatives with respect to the orientational Euler angles requires differentiating the Wigner rotation matrices. For alpha and gamma this accomplished by a straightforward application of the derivative of an exponential

$$\frac{\partial D^2_{m,n}}{\partial \alpha} = -ime^{-im\alpha} d^2_{m,n} e^{-in\gamma} = -imD^2_{m,n}$$

$$\frac{\partial D^2_{m,n}}{\partial \gamma} = -ine^{-im\alpha} d^2_{m,n} e^{-in\gamma} = -inD^2_{m,n} \tag{15}$$

but the derivative with respect to beta requires differentiating each element of the Wigner d-matrix. Using hose derivatives, given in table 2, we find the derivative of the elements of the Wigner rotation matrix with respect to beta to be

$$\frac{\partial D^2_{m,n}}{\partial \beta} = e^{-im\alpha} \frac{\partial d^2_{m,n}}{\partial \beta} e^{-in\gamma} \tag{16}$$

Substituting the Wigner matrix derivatives back into the derivative of the dipolar coupling we can write out the derivative of the dipolar coupling with respect to the bond-crystal orientation angles as

$$\frac{\partial\, \omega^{(1)}_{I_i,S_j}}{\partial \beta_{CB}} = -\frac{2\pi}{\sqrt{3}}\, a\, b_{IS} \sum_{\substack{m=-2 \\ m\neq 0}}^{2} D^2_{-1,-m}\left(\alpha_{RC}, \beta_{RC}, \gamma_{RC}\right) \frac{\partial\, D^2_{-m,0}\left(0, \beta_{CB}, \gamma_{CB}\right)}{\partial \beta_{CB}} \tag{17}$$

Further simplification is unproductive and a combined equation is unnecessary as the individual derivatives derived above can be individually computed and recombined to determine the final derivatives of the signal with respect to the parameters described by equation 9. The error in the parameter estimates is then calculated as

$$\sigma_p = \frac{dp}{dS} \times \sigma_y$$
$$p = a, \alpha, \beta, \gamma, \Gamma_2 ... \tag{18}$$

Results

We implemented the simulation model above in python using C++ and the Eigen linear algebra library [Guennebaud, 2010]. The simulation procedure is exposed to python as a compiled module and a convenient interface is implemented using the Numpy numerical library. An implementation of Differential evolution from the SciPy package [van der Walt, 2011] is used for rough fitting using Latin hypercube initialization to ensure the candidate population is uniformly distributed through parameter space. A randomized gradient descent optimization used for polishing is implemented in python with Numpy linear algebra functions.

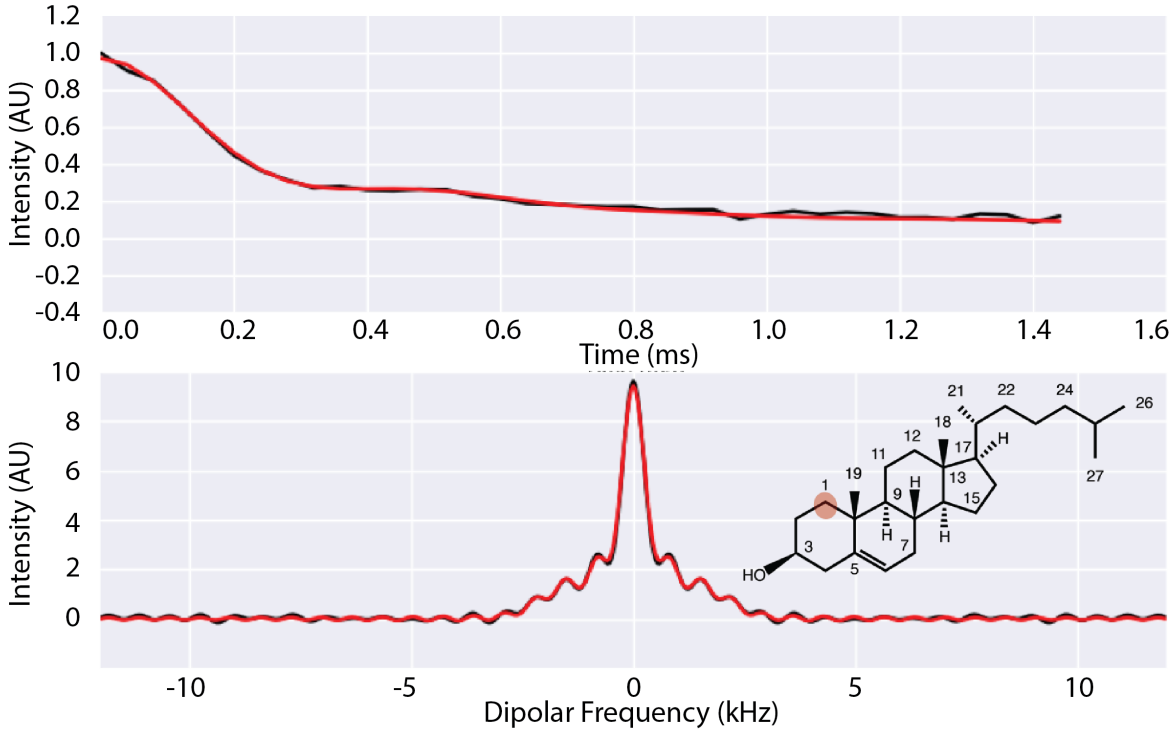

Figure 3.1 – ^{1}H-dephased ^{13}C intensity for carbon 1 of cholesterol measured with $R18_1^7$ recoupling sequence. Experimental data (black) Fit line (red). Fit using 3 protons to an RMSD of 0.017. Fit scaling factors are 0.110, 0.081, and 0.031.

A typical measured dipolar trajectory for a ^{13}C dephased by dipolar couplings to ^{1}Hs in a sample of uniformly ^{13}C-labelled cholesterol is shown in in Figure 3.1. This trajectory can be fit assuming only a single proton. As the bond frame is defined by a single interatomic vector and is rotationally symmetric around those bonds, the line shape is independent of the bond-crystallite Euler angles and only the scaling factor and only the relaxation rate and the single scaling factor are fit. The ability to fit with a single proton was determined by an increase in the reduced chi-squared statistic upon the inclusion of a second proton.

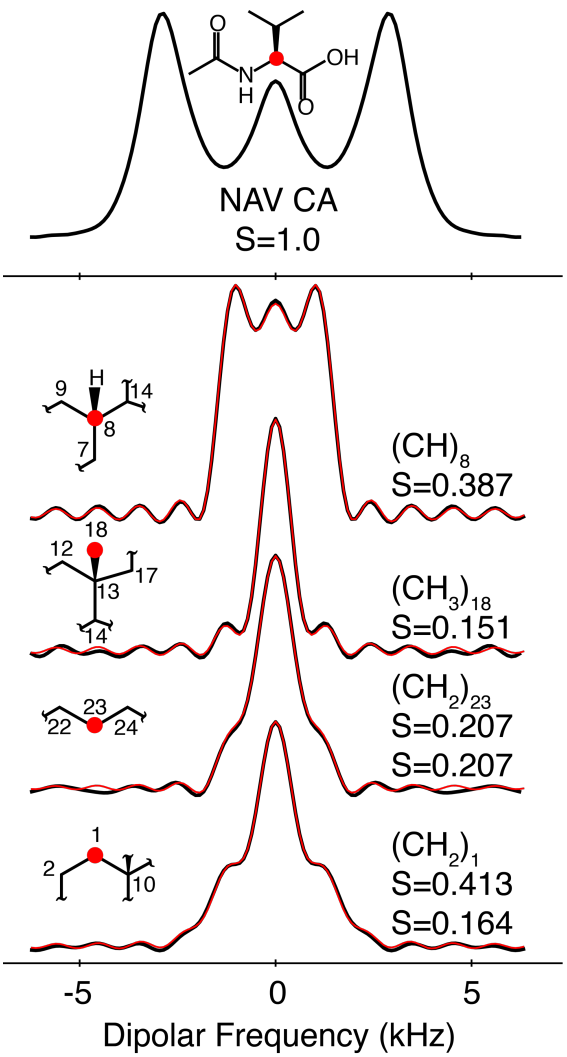

Figure 3.2 - A variety of different line shapes occur for different proton geometries and order parameters. N-acetyl valine (NAV) has large dipolar couplings due to high rigidity and little averaging. Other atoms throughout cholesterol (bottom four line shapes) have varying order parameters.

The differential evolution fitting procedure with C++ simulation implementation easily scales to higher proton numbers and high quality fits can achieved for a variety of proton environments, shown in the frequency domain (Fig. 3.2). To assess the exploration of the available parameter space, the parameters used in all the simulation function evaluations were logged and are displayed in Figure 3.3. We observed good coverage of the parameter space for every pair of parameters and saw no indication of "blind spots" where higher order combinations of parameters were not explored. Multiple fitting runs with different random seeds consistently found the

same general minimum which was polished by randomized gradient descent.

The major advantage of the fitting procedure presented here is the ability to vary every parameter independently and include an arbitrary number of protons. Figure 3.4 shows a pair of fits for the same dipolar trajectory, both using two protons, but the first with the scaling factors

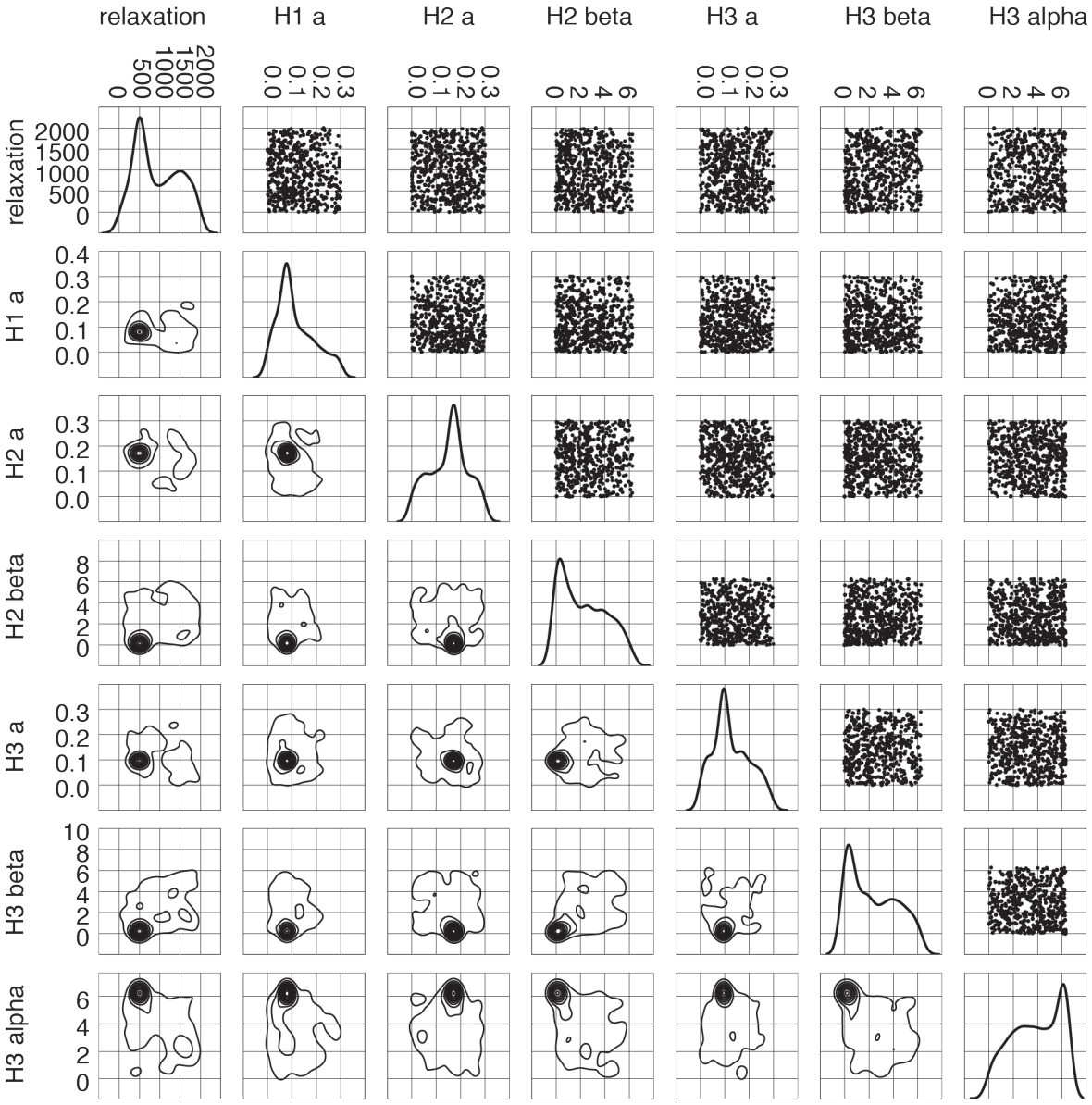

Figure 3.3 - Parameter search by differential evolution for a 3-proton fit. All values searched over during fitting procedure are displayed. Top right shows scatter plots for pairs of parameters. The diagonal shows a kernel density estimated distribution of the explored parameters. Bottom left shows 2D kernel density estimated distribution for pairs of parameters. All parameter pairs show good exploration without bounds and peaks where good parameters were identified.

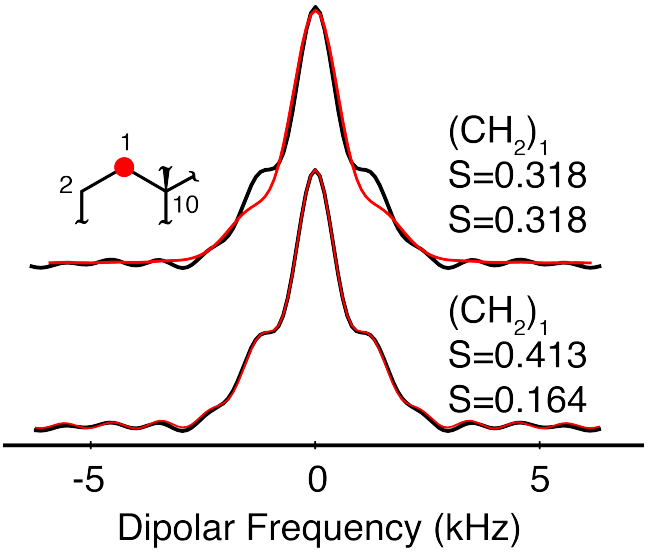

Figure 3.4 - Two fits for the same spin cluster around carbon 1 in cholesterol show very different qualities. The top fit has the order parameters of the two protons constrained ot be equal. A good quality fit can only be attained if the two are allowed to take different values (bottom).

tied. This carbon is a methylene and there is no indication that the bond lengths are significantly different. The poor fit with tied scaling factors and the excellent fit with independent scaling factors reveals a differential averaging of the dipolar couplings, presumably due to orientational differences of the two interatomic vectors with respect to the motional mode causing averaging. Similar trends of strongly varying order parameters are observed for many carbons in cholesterol and their analysis and interpretation will be detailed in an upcoming paper describing work with Lisa Della Ripa, Zoe Petros, and others.

To perform error analysis for a given fit, the derivatives only need to be calculated once to produce a variance-covariance matrix for the parameters. The variance of the parameters is given along the diagonal of the matrix and their correlation can be computed by normalizing the variance-covariance matrix to have all ones along the diagonal forming the correlation matrix (Fig. 3.5). Correlations among parameters vary significantly as a function of parameter values which can cause significant difficulties for gradient-based optimization routines causing strong oscillations even if compensating techniques are used such as momentum and momentum decay.

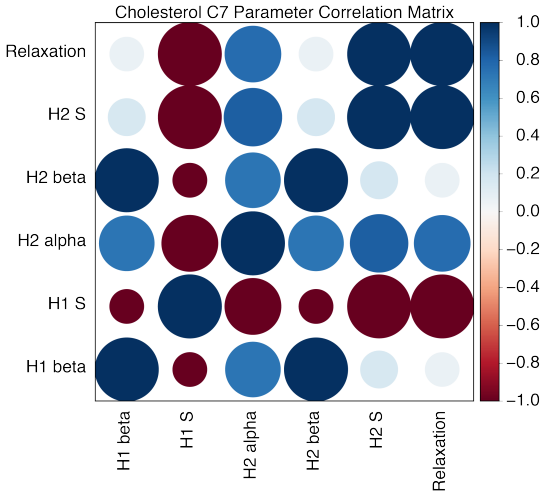

Figure 3.5 - Correlation matrix showing strong correlations between fit parameters for carbon 7 in cholesterol. Magnitude is represented by circle size and the value is represented by color.

Conclusions

Here we have described a simulation and fitting procedure for recoupled dipolar trajectories in the solid state that assumes no structural or dynamical models. We have shown significantly increase fit quality and the elucidation of detailed dynamical effects that were previously unattainable. We expect that robust and detailed fitting of dipolar trajectories will enable more detailed dynamical models of protein fibrils, membrane proteins, and other membrane components such as sterols and lipids which are difficult to study at atomic detail by other methods.

In further efforts, we plan to incorporate simulation of ^{13}C-^{13}C dipolar coupled transfer to enable cross correlation of X-^{1}H and X-X dipolar couplings to allow for detailed description of H-C-C-H and H-C-N-H dipolar coupled networks giving us access to structural details and dynamics

of important systems such as protein backbone and sidechains without imposing restrictive models and parameter equivalences enabling measurement of biochemical systems with unprecedented spatiotemporal detail.

References

- Klaus Schmidt-Rohr and Hans Wolfgang Spiess, Multidimensional Solid-State NMR and Polymers. Academic Press. New York. 1994.
- Ad Bax and Alexander Grishaev, Current Opinion in Structural Biology, 15:563–570 (2005)
- Malcom H. Levitt, Encyclopedia of Nuclear Magnetic Resonance, 9. 2002.
- M. Hohwy, C.P. Jaroniec, B. Reif, C.M. Rienstra, and R.G. Griffin, J. Am. Chem. Soc. 122, 3218-3219. 2000.
- Rienstra, C. M.; Hohwy, M.; Mueller, L. J.; Jaroniec, C. P.; Reif, B.; Griffin, R. G. J. Am. Chem. Soc. 2002, 124, 11908-11922.
- Christopher John Willy. Recursive Parameter Estimation of Markov-Modulated Poisson Process. PhD thesis, The George Washington University, 2008.
- Gael Guennebaud and Benoit Jacob et al., Eigen v3. http://eigen.tuxfamily.org, 2010.
- Stéfan van der Walt, S. Chris Colbert and Gaël Varoquaux. Computing in Science & Engineering, 13, 22-30 (2011),

CHAPTER 4: Global Peak Scoring: Peak Prediction for Constraint Identification and Model Validation

Introduction

NMR spectra benefit but also suffer from their sensitivity to wide variety of structural and dynamic details, which give the experimenter an ability to manipulate the effective Hamiltonians to choose specific features to emphasize. The experimental flexibility and richness of the data make NMR a valuable method for answering research problems, but it simultaneously makes modelling data and predicting measurements extremely difficult. For example, the NOE is a fundamental phenomenon used to determine thousands of protein structures. Yet only recently has it become possible to accurately predict the peak volume observed in NOESY spectra, and this calculation required the usage of high performance computing resources and a near complete characterization of all relevant system details [Edwards, 2014].

Because of its complexity and the great extent of knowledge required to judge spectral characteristics, NMR analysis has remained primarily manual and in the realm of experts. It is challenging to apply purely objective metrics in NMR analyses, and in particular the method of ranking a model's quality by direct comparison of simulated and experimental data—the most common approach in most physical sciences—is limited in NMR analyses to problems that can be accurately described by a two-spin model. For larger spin systems such as in proteins, cross-validation methods therefore are limited to the comparison of a structural model with experimental parameters, such as residual dipolar couplings (RDCs), that depend on only two spins to an excellent approximation [Bax 2003, Simon 2005]. However, these approaches are inherently limited to systems for which both RDC and other types of data can be acquired in sufficient quality and quantity to enable the structure calculation to converge by including only

subsets of the restraints. More generally, cross-validation of protein structures by solution NMR is limited to cases where NOE, Karplus and TALOS restraints are available in high quality, in addition to RDCs and/or residual chemical shift anisotropies .

Here we describe efforts towards a general framework for predicting spectra from given protein structures and empirical descriptions of coherence transfer mechanisms and their use in constraint identification and model validation. The aim of this work is to provide a unified metric for judging the internal consistency of an analysis and providing a robust test of model quality.

Methods

Peak Enumeration

Peaks in protein NMR correlation spectra arise from magnetization transfer pathways through the protein. It is possible to frame the problem of enumerating all the peaks expected in an k-dimensional spectrum as the search for all k-length paths in a graph (representing the

```
ncacx_primary = [
    ( # 1st atom specification
        None,  # No restriction of sequence position for first atom
        None,  # Residue type not restricted
        ['N'], # Atom type restricted to amide nitrogens
        None   # Restrictions on distance from previous atom not applicable
    ),
    ( # 2nd atom specification
        [0],   # Restricted to same residue
        None,
        ['CA'],# Restricted to alpha carbons
        None
    ),
    ( # 3rd atom specification
        [-2,-1,0,1,2], # Allow transfers to i+/-2 residues
        None,
        ['C.*'], # allow transfer to all carbons within allowed residues
        None
    )
]
```

Listing 4.1 – Example coherence pathway specification. The primary coherence transfer pathway for a moderate mixing N-Cα-Cx type experiment which is intended to emphasize intraresidue correlations but may also include neighboring residues as weaker peaks.

protein) that meet a set of criteria. The naive approach to the general problem of enumerating all k-length paths in a graph takes time that is proportional to the number of nodes raised to the power of k. Here we use a more efficient method which involves building a set of trees, one for each starting atom, where the children of each node are the atoms the coherence can travel to. This tree-based algorithm has a running time that is linear in the number of atoms and proportional to the product of the branching factors of each transfer step; but, for many coherence transfers, there is a very small number of neighboring atoms to which the coherence can transfer, which drastically limits computation time.

Each coherence transfer pathway of length k consists of k sets of restrictions on the neighbors to consider. In an effort to be as general as possible, the specification of a pathway has the same structure for every experiment type. For each transfer the atoms that can participate in the pathway can be restricted by five criteria:

- Position in amino acid sequence relative to previous dimension
- Residue type
- Atom type
- Number of bonds from previous atom
- Through-space distance from previous atom

Using these restrictions, the coherence transfer trees are constructed by a recursive, depth-first search (DFS) through the graph of all possible transfers through the protein.

As an example, the specification for the primary coherence transfer pathway of an NCACX specification is given in Listing 4.1. It demonstrates the use of the first three types of restriction. Additional examples in Listing 4.2 demonstrate the use of specifications to include

non-standard coherence transfer pathways and the specification of bond distance and through-space distance restrictions.

The algorithm (Listing 4.3) consists of four major operations: (1) enumerate valid atoms; (2) enumerate valid atom pairs involved in each transfer; (3) construct the tree, and (4) walk along the branches. Here we elaborate on each operation.

```
# A)
ncacx_proline = [
    (None, ['P'], ['N'], None),
    ([0], ['P'], ['CD'], None),
    ([-2,-1,0,1,2], None, ['C.*'], None)
]

# B)
ncacbco_primary = [
    (None, None, ['N'], None),
    ([0], None, ['CA'], None),
    # peaks are restricted to 0 or 1 bonds away
    ([0], None, ['C.*'], ('b',[0,1]))
]

# C)
long_cc_primary = [
    (None, None, ['C.*'], None),
    # atoms must be within 5 angstroms
    (None, None, ['C.*'], ('s', [5.0]))
]
```

Listing 4.2 - Additional example coherence pathway specifications. A) coherence pathway for N-Cδ-Cx in proline during NCACX type experiment. B) NCACBCO coherence pathway demonstrating the restriction of the Cα-Cβ/C' transfer to distances of exactly 0 or 1 bond allowing N-Cα-Cα, N-Cα-Cβ, and N-Cα-C' peaks. C) coherence pathway for long-mixing through-space ^{13}C-^{13}C expderiment demonstrating a through-space distance restriction for the second dimension.

(1) Enumeration of valid atoms is a simple filtering procedure. For each transfer step, the function enumerate_valid_atoms loops through each atom in the protein, determines if it complies with the restrictions, and if so, adds it to the set of valid atoms for that transfer.

(2) Similarly, the function enumerate_valid_transfers performs a filtering procedure by looping through each pair of atoms in the protein.

(3) The function build_trees is passed a tree node, the parent atom it corresponds to, the lists of valid atoms and transfers, and the current depth. If the current depth is equal to the maximum depth, the function returns the current tree node without adding any. Otherwise it loops through all valid transfers at that depth that start on the parent atom and adds children to the parent node by calling the build_trees function itself with the child atom as a new parent atom and an empty tree node to store its children in. The function is initially called without a parent atom and this node has all valid atoms in the first dimension as its children.

(4) The collected pathways are harvested from the tree by a depth-first search that emits the full transfer pathway taken once it arrives at a terminal node. The pathways are then represented as k-tuples of atoms. To convert those pathways into predicted peaks, the steps in the pathway not corresponding to observed dimensions are removed and the remaining pathways are translated into their chemical shifts, if known. The resulting peak list contains every peak that is expected to exist in a spectrum for a given structural model of the protein under the assumptions that (a) the set of allowed pathways is complete for the pulse sequence, (b) the resonance assignments are correct.

```
function enumerate_valid_atoms(sequence, dim_spec):
    set valid_atoms = {}
    for residue in sequence:
        if residue.type in dim_spec.allowed_residue_types:
            for atom in residue:
                if atom.type in dim_spec.allowed_atom_types:
                    valid_atoms.add(atom)
    return valid_atoms

function enumerate_valid_transfers(valid_atoms_1, valid_atoms_2, dim_spec):
    set valid_atom_pairs = {}
    for atom_1 in valid_atoms_dim_1:
        for atom_2 in valid_atoms_dim_2:
            delta_idx = (atom_2.residue.index - atom_1.residue.index)
            if delta_idx in dim_spec.allowed_relative_sequence_positio and\
            bond_distance(atom_1, atom_2) in dim_spec.allowed_bond_dist and\
            distance(atom_1, atom_2) in dim_spec.allowed_spatial_dist:
                valid_atom_pairs.add((atom_1, atom_2))
    return valid_atom_pairs

function build_trees(
    valid_atoms,
    valid_transfers,
    current_depth = 0,
    tree = Tree(),
    previous_atom
):
    if current_depth == len(valid_atoms):
        return tree

    for (atom_1, atom_2) in valid_transfers[current_depth]:
        if atom_1 == previous_atom and atom_2 in valid_atoms[current_depth]:
            tree[atom_2] = build_trees(
                valid_atoms,
                valid_transfers,
                tree = Tree(),
                current_depth = current_depth + 1
                previous_atom = atom_2
            )
    return tree

function enumerate_transfer_pathways(sequence, pathway_spec):
    list valid_atoms = [[],[],[]]
    for i in range(pathway_speci.num_dimensions):
        valid_atoms[i] = enumerate_valid_atoms(sequence, pathway_spec [i])

    list valid_transfers = []
    for j in range(pathway_spec.num_dimensions - 1):
        valid_ transfers[j] = enumerate_valid_transfers(
            valid_atoms[j],
            valid_atoms[j+1],
            pathway_spec[j+1]
        )
    tree = build_trees(valid_atoms, valid_transfers)
```

Listing 4.3 - Functions implementing described algorithm written in Python-esque pseudocode.

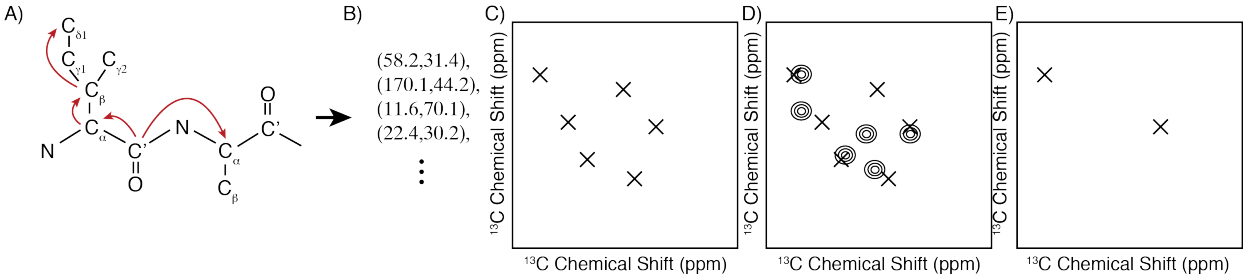

Figure 4.1 – Flow chart for an algorithm to filter peaks with experimental data. A) predicted coherence transfers are enumerated using the tree-based algorithm described in the text. B) the coherence transfers are converted to peaks by usage of assigned resonances. C-E) The peaks are compared against experimental data and those that overlap portions of the spectrum with low signal intensity are interpreted to be incorrect predictions and are discarded leaving only those peaks that are consistent with the data.

Peak Filtering

The enumerated predicted peaks are effectively a hypothesis in need of testing. To test that hypothesis, the predicted peaks are compared against experimental data. Given a spectrum of the type for which peaks were predicted, intensity values are sampled at the location of each peak. If the intensity at the peak's location is greater than a given threshold, the peak can be assumed to exist in the data. Whether the peak arose from the coherence pathway predicted is still unknown, because real data contains overlapping peaks, noise, and artifacts. However, if the intensity exists in the data at the location of the peak, one must assume that that coherence pathway could contribute to the observed data. A diagram of this process is shown in Figure 4.1.

Filtering the predicted peaks through experimental spectra offers a way to measure what portion of the spectrum is explained by the protein model and coherence transfer pathway. The filtered peaks can also be used as preliminary assignments for the identified peaks. Because the process does not discriminate between the likelihood of pathways the resulting peak list represents an objective prediction of the assignments of peaks in a spectrum.

Ambiguous Distance Assignments

Once a set of peaks has been identified as possibly existing in the data, grouping them based on proximity can be used as a proxy for predicting overlap in the experimental spectrum. In doing so, peaks can be consolidated into ambiguous assignments of spectral intensity. For through-space mixing experiments such as long-mixing dipolar assisted rotary resonance (DARR) and proton-assisted recoupling (PAR) experiments, this facilitates the possibly ambiguous assignment of long-range distances in proteins. The resulting groupings can be returned by a script as a list of constraints with distance bounds based on spectral intensity for use in structure determination software.

Results and Discussion

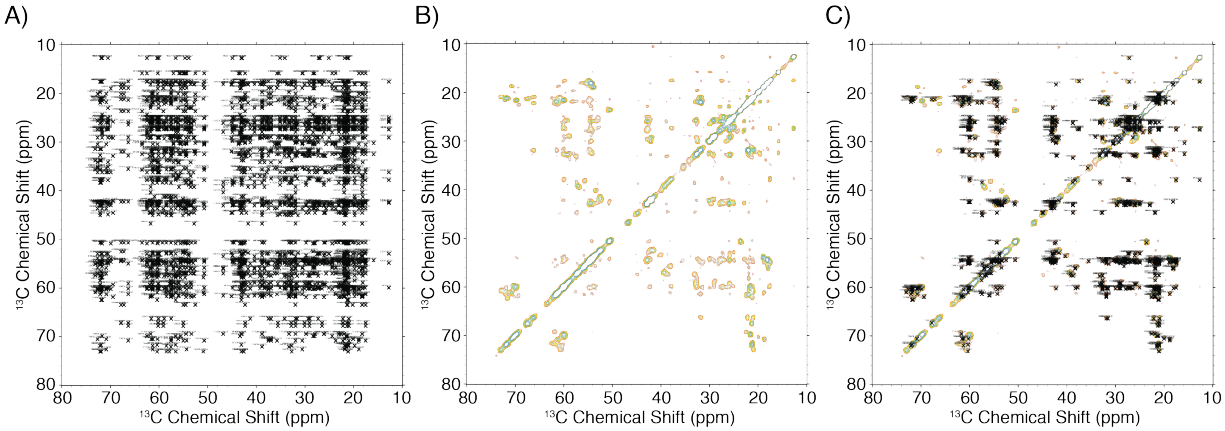

Figure 4.2 - Demonstration of peak filtering on GB1 PAR spectrum. A) Predicted peaks are enumerated in the aliphatic region of the spectrum. B) An example experimental ^{13}C-^{13}C PAR spectrum. C) Peaks from A) filtered to only include those located in areas of intensity above a signal-to-noise ratio of 6.

A demonstration of the algorithm in Figure 4.1 is shown for a ^{13}C-^{13}C GB1 PAR spectrum (for which through-space ^{13}C-^{13}C polarization transfers occur, mediated by intervening protons). In Figure 4.2a the peaks predicted using a through-space coherence pathway are displayed with a distance cutoff of 8.5Å as measured in PDB entry 2LGI. The predicted peaks are densely covering the entire aliphatic region. When overlaid with a spectrum of the same type

predicted (Fig. 4.2b) and filtered to only include peaks where data is present (Fig. 4.2c) the surviving number of peaks is much lower indicating that either the protein structure used in predicting the peaks was incorrect, the distance cutoff in the coherence pathway specification was inappropriate, or the description of the coherence pathway is inadequate in some other way.

It is common practice to assume that spectra collected for the purpose of measuring distance restraints will always contain short-range contacts, but that long-range contacts can appear somewhat stochastically. However, coherence transfer mechanisms often depend on the inverse sixth power of the distance, which can confound prediction of peak patterns because dynamic averaging of the r^{-6} factor heavily skews the weighted average towards representing the short end of the range of motion. For this reason, it is usually best to assume that long-range distances may or may not appear and that weak peaks may represent very long-range contacts on average. Therefore, it is often advantageous to predict peaks out to the furthest expected average distance that could produce peaks and ignore many of the unobserved peaks. Specifically this means that for analysis of through-space mixing spectra, it is most useful to predict peaks with a large distance cutoff and expect some fraction of incorrect predictions. Therefore it is somewhat expected that the static coherence transfer pathway can only describe 1562 out of 7931 predicted peaks.

Another type of spectrum used for assignment purposes is the COSY spectrum which relies on through-bond J-coupled transfers. In this type of experiment, only carbons one bond away from each other are expected to produce signals. Applying the algorithm in Figure 4.1 to a COSY spectrum of GB1 results in 100% of the predicted peaks being observed in an experimental spectrum indicating that the coherence pathway specification and the protein sequence used in peak prediction are completely adequate to explain the data.

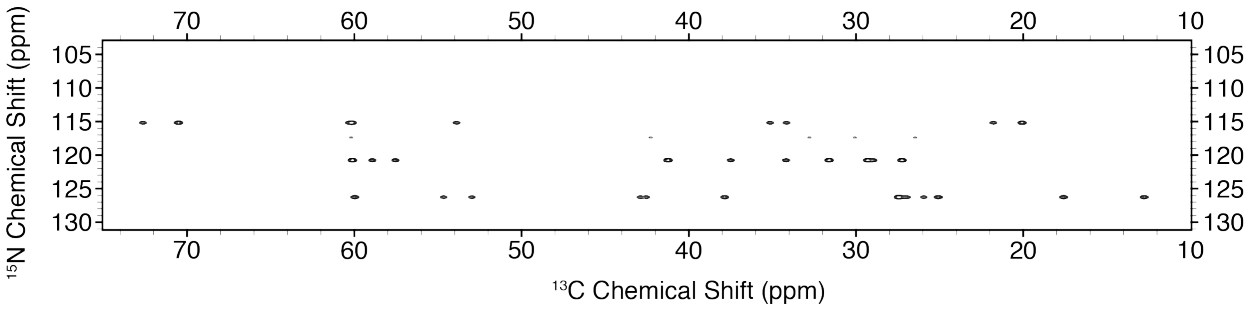

Figure 4.3 – A simulated plane from a ^{15}N-^{13}Ca-^{13}Cx spectrum of GB1 with moderate mixing. Plane shown is centered at 60 ppm. Peak intensities are chosen to be 100 for intraresidue peaks and 50 for neighboring interresidue peaks.

One possible further applications of the GPS approach is as a final check of self consistency of the analysis of a spectral dataset. Since the GPS-predicted peak lists rely sensitively on the structural model and experiment description, comparing them against experimental data is a strong test of the internal consistency of an analysis. A simple possibility is to count the portion of predicted peaks observed in the data. For instance, we could score the structure and PAR coherence pathway specification used above with a 1562/7931 or approximately 19.7% accuracy indicating poor agreement (in this case due to an overestimation of distance cutoff). Combined metrics that aggregate the explained peaks in all spectra in a dataset would be an overall score of the portion of the data explained by the model. Such a metric is thus far missing in the NMR literature and would be extremely valuable, acting as a goal post for analyses.

A more sophisticated method for testing internal consistency of an analysis could incorporate the prediction of spectral intensities as well as peak positions. Shown in Figure 4.3, a predicted plane of an ^{15}N-^{13}Cα-^{13}Cx contains intraresidue peaks and peaks from nearest neighbors. The intensities are arbitrarily assigned values of 100 for intraresidue correlations and 50 for nearest-neighbors. While this intensity prediction is somewhat arbitrary, it captures the

qualitative patterns observed in moderate-mixing ^{15}N-$^{13}C\alpha$-^{13}Cx spectra and represents a possible further method for testing predictions. If a suitable comparison metric can be determined for comparing spectral intensity directly, the abstraction and simplification of spectra to peak positions could be avoided.

Conclusions

Here we have presented a unified scheme for predicting peaks in a spectrum and using them for spectral analysis. We have proposed a possible method for checking the combined analysis of a spectral dataset for self-consistency that we expect to facilitate more reproducible analyses. We aim to further expand upon these scoring methods to enable a robust and universal score for describing the agreement between model and data that could be used as a general metric for the quality of a model.

References

- Luke J. Edwards, D.V. Savostyanov, Z.T. Welderufael, Donghan Lee, Ilya Kuprov. Journal of Magnetic Resonance. 243, 107-113. 2014.
- Ad Bax. Protein Science, 12, 1-16. 2003.
- Katya Simon, Jun Xu, Chinpal Kim, and Nikolai R. Skrynnikov. Journal of Biomolecular NMR. 33, 83-93. 2005.

CHAPTER 5: High-Resolution Structure Refinement of Human α-Synuclein Fibrils with Proton Distance Restraints

Work with Marcus D. Tuttle, Andrew J. Nieuwkoop, and Chad M. Rienstra

Introduction

α-Synuclein (α-syn) fibrils, the primary constituent of Lewy bodies and Lewy neurites, have been implicated in the propagation of PD-like pathology via neuron-to-neuron transfer. Indeed, inoculation of wild-type (WT) mice with preformed α-syn fibrils leads to the recruitment of endogenous α-syn into intracytoplasmic inclusions and the reproduction of many features of the

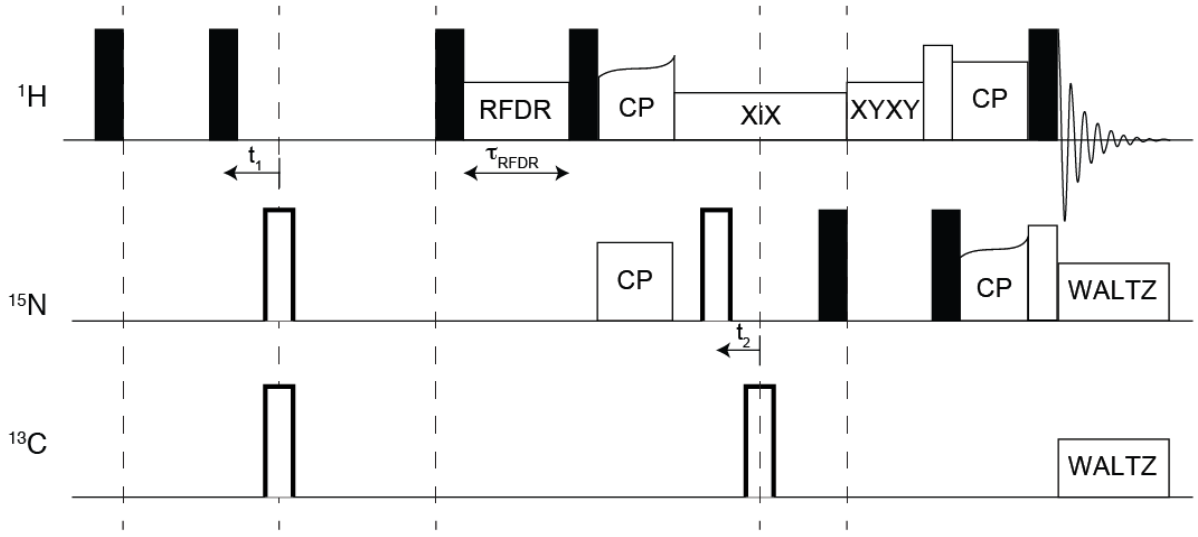

Figure 5.1 - Three-dimensional, ^{1}H-detected pulse sequence for measuring ^{1}H-^{1}H distances in proteins. A) HhNH 3D pulse sequence. Filled and empty rectangles represent pi/2 and pi pulses respectively. States-TPPI was applied to the initial pi/2 pulse on ^{1}H and the ^{15}N cross polarization field before F2 acquisition B) Residual water signal is now well separated from water-protein correlations and potential amide-aliphatic correlations.

neurodegenerative cascade. [Luk, 2009] Recently we presented the first structure of a pathogenic α-syn fibril, which utilized extensive ^{13}C-^{13}C and ^{15}N-^{13}C distances derived from 2D and 3D ^{13}C-detected magic-angle spinning (MAS) solid-state NMR (SSNMR) experiments. In that study, spectra were analyzed manually and structures calculated with XPLOR-NIH using standard protocols for utilization of ambiguous restraints.[Schwieters, 2006] The resulting structure was

validated by electron microscopy and X-ray fiber diffraction and is a conserved structural fold for at least two of the early onset PD mutants (A30P and A53T).[Lemkau, 2012] This first α-syn structure was computed with thousands of cross peaks observed in 68 spectra. Similarly, structures of amyloid-β (Aβ) (1-40), Aβ(1-42), relied almost entirely on ^{13}C-^{13}C, ^{15}N-^{15}N, and ^{15}N-^{13}C distances determined from a variety of pulse sequences (DARR, PAR, PAIN, REDOR and TEDOR). [Lu 2013,Wälti, 2016]. These structures based on ^{13}C-detected experiments were possible because of the large chemical shift dispersion of ^{13}C, the availability of sparse labeling patterns (derived from 1,3-^{13}C-glycerol, 2-^{13}C-glycerol, 1-^{13}C-glucose or 1,6-^{13}C-glucose) to achieve line narrowing, high magnetic field and/or site-specific labeling by solid-phase peptide synthesis.

Clearly it would be beneficial to develop and apply improved methods that can increase speed and sensitivity while decreasing required sample quantities and effort required both for data collection and analysis.

Recently there has been a flourishing of proton detection methodologies that leverage fast MAS instrumentation and newly developed pulse sequences. Early applications of these approaches were limited to small, crystalline proteins such as SH3 and GB1 at high levels of deuteration, [Akbey, 2010, Zhou, 2012].

Spectra of amyloids have been collected (^{1}H-^{15}N), but thus far not used for structure determination. Therefore, ^{1}H-^{1}H distance measurements present an attractive opportunity to measure a largely orthogonal set of restraints on fibril structure that gives access to many intermolecular contacts that are not apparent in ^{13}C-^{13}C and ^{15}N-^{13}C mixing data. These advantages are amplified in the case of small amino acids with few carbon and nitrogen atoms which are especially abundant in amyloidogenic proteins.

To collect ^{1}H-^{1}H correlation data we prepared uniformly-^{2}H,^{13}C,^{15}N labeled (DCN) α-syn which was back-exchanged ^{1}H at all exchangeable hydrogen sites prior to fibrillization. We then collected data at 36 kHz MAS using a modified version of the common HNH experiment shown in Figure 5.1a. Here we modified the pulse sequence to perform the indirect ^{1}H dimension frequency labeling prior to ^{1}H-^{15}N cross-polarization. This approach, when combined with MISSISSIPPI, [Zhou, 2012], greatly improves the dynamic range and sensitivity of the 3D spectrum by virtue of improved water suppression. We attribute this to the minimal period between solvent suppression and detection, which minimizes the potential for longitudinal relaxation of the water signal. This approach is particularly helpful for resolving the Hα and other sidechain peaks that are present due to residual protonation of the non-exchangeable sites in the bacterial expression. The better separation of residual water signal in the direct dimension from real protein-water correlations and potential amide-Hα correlations is shown in part B of Figure 5.1.

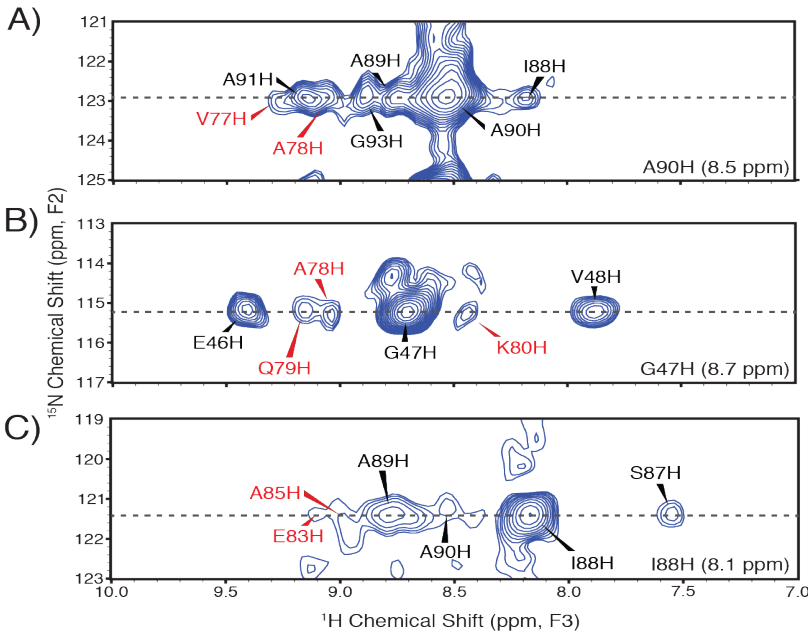

Figure 5.2 - 2D planes from HhNH 3D experiment collected on 100% ^{1}H back-exchanged CDN α-synuclein with 7.2 ms RFDR mixing at 36 kHz mas rate. Black labels indicate near-neighbors while red are long-range correlations. A) 2D plane at A90 amide proton frequency (8.5 ppm). B) 2D plane at G47 amide proton frequency (8.7 ppm). C) 2D plane at I88 amide proton frequency (8.1 ppm).

As we show in Figure 5.2, some long-range distances could be unambiguously assigned from the 3D spectra. The majority of ^{1}H signals, however, were partially ambiguous due to the

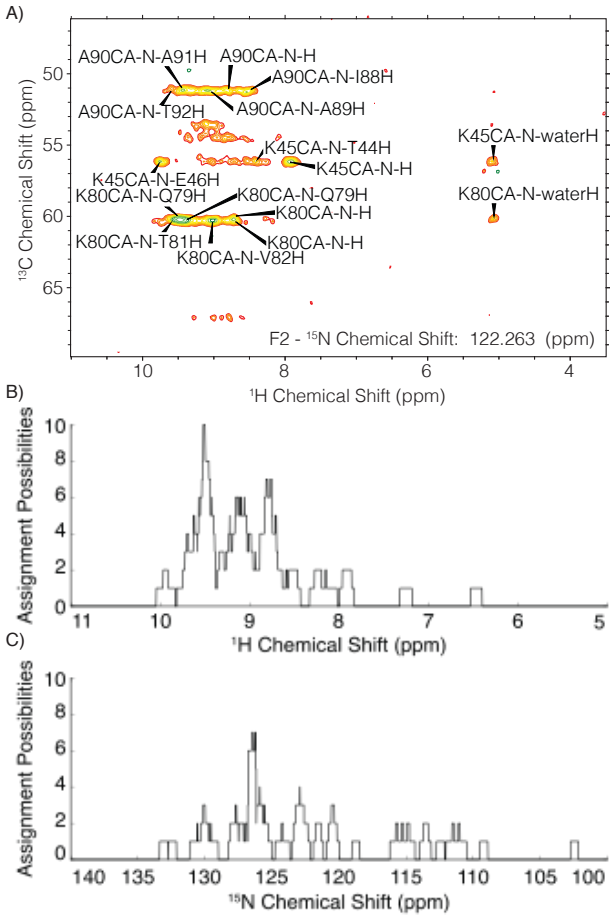

Figure 5.3 - Number of possible assignments for a peak observed at a given frequency based on the average linewidth in that dimension (0.2 ppm for 1H and 0.65 ppm for 15N) for the HhNH 3D shown in Fig. 1 and 1H and 15N assignments for α-synuclein reported in BMRB entry 18243 A) Ambiguity of assignment as a function of 1H chemical shift. B) Ambiguity of assignment as a function of 15N chemical shift. Distributions produced using kernel density estimation with tophat kernel.

preponderance of amide shifts in the range of 8.5 to 10 ppm and typical linewidths of 0.1 ppm. A typical plane can be observed in Figure 5.3a where correlations to A90Cα-N and K80Cα-N are severely overlapped in the amide proton region of the direct dimension. This results in two- to five-fold assignment ambiguity for most amide protons (Fig. 3). Likewise, there remains some ambiguity in the direct ^{15}N dimension, given the observed 0.5 ppm linewidths for ~60 signals dispersed over 20 ppm. Thus the indirect ^{1}H dimension of observed peaks have up to 10 possible *a priori* assignments and the ^{15}N chemical shifts have up to 7. We identified 27 unambiguous, long-range distances but were left without assignments for many observed correlations.

A number of approaches are available for analyzing ambiguous distance restraints. However, most available approaches use an iterative algorithm that repeatedly compares a reassigns correlations by comparison with a structure computed from the previous iteration of assignments.

This approach can bias calculated structures towards incorrect initial assignments leading to convergence in suboptimal minima. Though there are approaches to address this problem we chose to avoid it entirely by not comparing a structure with the data used to compute it initially. [Schwieters, 2006, Linge, 2003]

To address the ambiguity observed in our data, we applied a semi-automated analysis protocol that leverages our previous structure determined using only ^{13}C-^{13}C and ^{15}N-^{13}C constraints to reduce assignment possibilities to a level that can be adequately disambiguated by mutual consistency in simulated annealing structure calculations.

Using the correlation enumeration algorithm of the Global Peak Scoring method, we enumerated all possible ^{1}H-^{1}H correlations within inter-proton distance cutoffs that would be considered generous for the given mixing times to address uncertainty in the initial structural ensemble. A HhNH coherence transfer pathway specification for GPS is given in Listing 1. Previously determined chemical shift assignments are then used to construct a list of possible peaks given the inter-proton distances observed in the initial structure. Upon comparison of the possible peak list to the experimental data we removed any peaks for which the signal was below five times the root-mean-square noise level. At this point the filtered peak list contains all correlations that may be

```
hhnh_primary = [
    # no restrictions on initial proton
    (None, None, ['H.*'], None),
    # proton-proton transfer limited to 5 angstroms
    (None, None, ['H.*'], ('s', [5.0])),
    # cross polarization to directly attached nitrogen
    ([0], None, ['N.*'], ('b',[1]))
    # cross polarization back to directly attached proton
    ([0], None, ['H'], ('b',[1])),
```

Listing 1 – Coherence pathway transfer specification for HhNH experiment

observed in the data regardless of mutual compatibility. We then further limited the peak list to those peaks with assignment ambiguity of less than 4. That is, we remove all peaks that are within a linewidth of more than three other predicted peaks. This filtering procedure left us with 86 new unambiguous ^{1}H-^{1}H distances and 66 new ambiguous ^{1}H-^{1}H distances.

To ensure a complete enumeration of possible distance constraints we applied the same structure-guided assignment procedure to the ^{13}C-^{13}C correlation spectra used in our previous structure calculation and despite extensive previous manual analysis of the data we determined an additional 2690 new unambiguous and 2246 new ambiguous ^{13}C-^{13}C constraints, which reinforced previously assigned correlations.

These ambiguous ^{1}H-^{1}H and ^{13}C-^{13}C constraints were then incorporated into XPLOR-NIH simulated annealing structure calculations by combining them with the previous structural restraints and pseudopotentials used to obtain the first α-syn structure we previously presented [Tuttle, 2016]. As these new restraints are from undiluted isotopically labeled samples, they were added to the pseudopotential that does not assume that individual pairs of nuclei give rise to observed correlations, and instead averages over the 10 monomers present in the structure calculation, as discussed in Tuttle *et al.*. The new ambiguous restraints were incorporated using the maximum inter-proton distance used in the automated analysis procedure as the upper distance

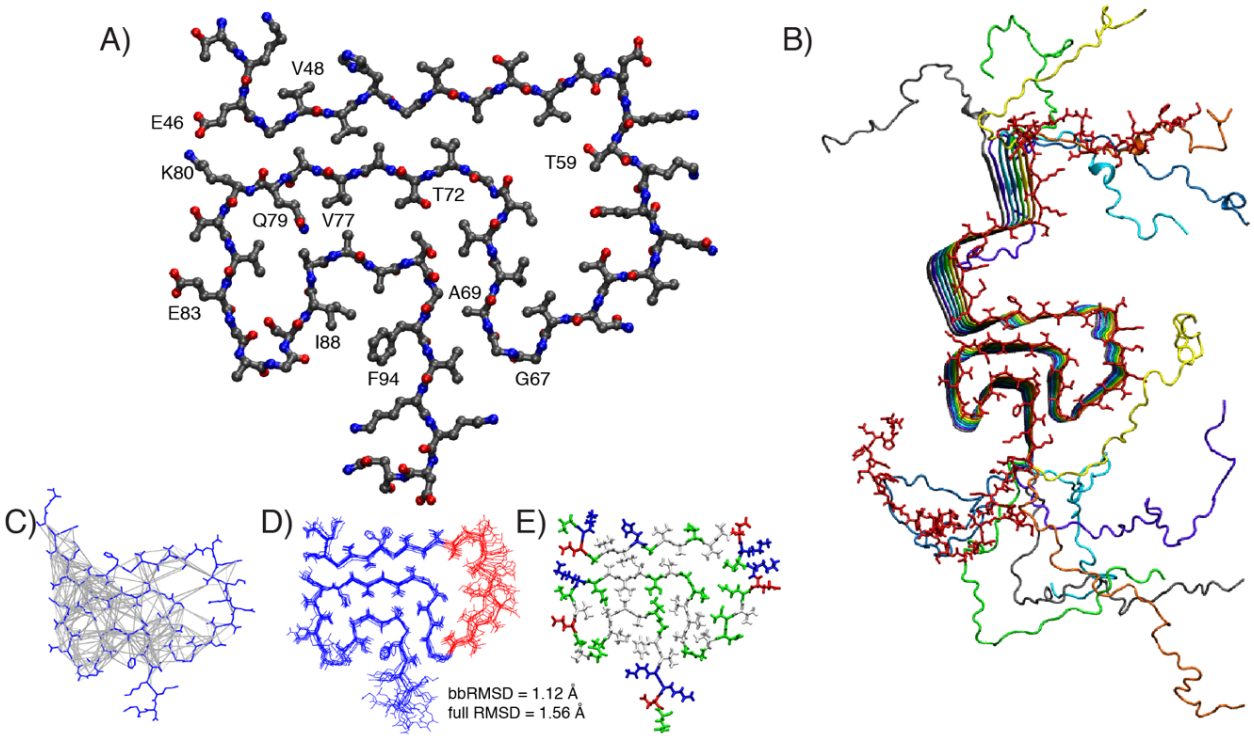

Figure 5.4 - Refined structure of α-synuclein fibrils utilizing ambiguously assigned ^{1}H and ^{13}C distances. A) Core of the central monomer of the lowest-energy fibril structure from residue 44 to 99. B) Full length structure of the 8 central monomers of the lowest-energy fibril structure. C) Map of all new, ambiguous ^{1}H and ^{13}C restraints used in the refinement. D) Overlay of the central monomer from the 10 lowest energy structures aligned using the backbone atoms for the structured residues. Blue indicates structured residues while red indicates the disordered loop from residue 55 to 66. E) Core structure colored according to hydrophobicity. White indicates hydrophobic residues, green, hydrophilic, red and blue, positively and negatively charged residues respectively.

limit of the restraint. These calculations converged to a single backbone fold consistent with the previously reported structure of α-syn fibrils.

The resulting structure, shown in Figure 5.4 shows substantially improved convergence of both the backbone and side chain atom positions. All new ambiguous ^{1}H-^{1}H and ^{13}C-^{13}C constraints are indicated in Figure 5.3c as gray lines. The improved determination of side chain atom positions

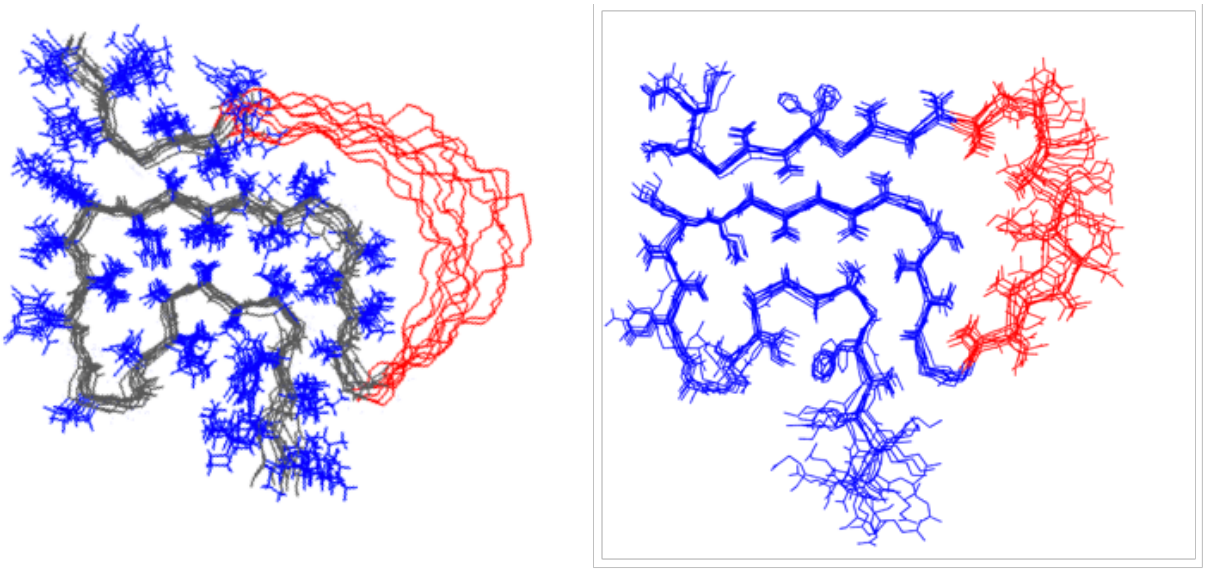

Figure 5.5 - Direct comparison of monomer structure before and after refinement. Left, structure from Tuttle *et al.* before refinement showing ensemble of 10 structures. Right, structure after refinement showing ensemble of 10 structures. RMSD calculations are performed between blue residues that are in common between two models.

throughout the core results in well-defined rotameric states for the majority of residues as can be seen in part D. The inclusion of new constraints has had an especially large impact on convergence of the positions of small amino acids including residues V52, A53, T54, V55, and A56 which now extend the N-terminal beta sheet into the disordered loop. The ^{1}H-^{1}H constraints have the largest impact on the convergence of the disordered loop around T59 which now exhibits defined preferences for the charged side chains to point to the exterior of the fibril while T59 and Q61 point inwards. The orientation of these hydrophilic side chains towards the previously unsatisfied T72 side chain provide greater definition to what could is likely a hydrogen bonding network or a pocket of water inside the disordered loop.

In conclusion, we have developed an objective, semi-automated procedure for refining SSNMR protein structures using 1H-detected spectra with suboptimal peak separation and have used it to refine the structure of a pathogenic fibril of α-syn. Using that previously untapped source of structural information for amyloid fibrils we have increased the resolution from to 2.04 Å to 1.56 Å.

We envision that refinement strategies of this type could be leveraged for study of other amyloid systems, and further improvements in the analysis algorithms are likely.

References

- Luk, K.C. et al. Exogenous α-synuclein fibrils seed the formation of Lewy body-like intracellular inclusions in cultured cells. Proc. Natl. Acad. Sci. USA 106,
- 20051–20056 (2009).
- Schwieters, C.D., Kuszewski, J.J. & Clore, G.M. Prog. Nucl. Magn. Reson. Spectrosc. 48, 47–62, 2006.
- Lemkau, L.R. et al. J. Biol. Chem. 287, 11526–11532, 2012.
- Marielle Aulikki Wälti, et al. Proc. Nat. Acad. Sci. U.S.A., 113, 34. E4976-E4984, 2016.
- Jun-Xia Lu et al. Cell 154, 1257-1268, 2013.
- U Akbey et al. J Biomol NMR. 46, 1, 67-73. 2010.
- Donghua Zhou et al. J. Biomol. NMR, 54, 291-305, 2012.
- Jens P. Linge Michael Habeck Wolfgang Rieping Michael Nilges. Bioinformatics 19, 2. 315-316. 2003.
- Marcus Tuttle et al. Nat. Struct. Mol. Biol. 23, 5, 409-415. 2016.

CHAPTER 6: Pathologically Distinct α-Synuclein Fibril Strains That Share A Common Tertiary Structure

Introduction

The structure and folding of protein fibrils is quite distinct from typical protein folding. The structure of pathogenic protein fibrils is unlikely to be an evolutionarily convergent structure. In fact, it has been proposed that any protein can misfold into a fibrillar state [Dobson reference]. Because fibrillary structures are "accidental," it is reasonable to postulate that they are not uniquely stable and that fibrillar misfolding in amyloid diseases may take many forms, a feature known as polymorphism. The ability of a fibril to elongate by the recruitment of flexible monomer in a templated fashion allows possibly metastable fibril forms to propagate, and as they lengthen, the transition into another fibril form becomes increasingly energetically disfavored; the initiation of a new fibril form would have to occur by a new seeding event or by cross-seeding. Templated replication allows fibrils to propagate as strains retaining the structure of their progenitors with good fidelity. The seeding of fibrils is not well understood mechanistically but is known to be a slow process, much slower than fibril propagation, presumably leading to the domination of the kinetically favored fibril form *in vivo*. Because of these features, fibril polymorphism has been hypothesized to be a cause of differential disease progression in Alzheimer's disease [Tycko reference] and possibly others.

The possible presence of fibril polymorphism in parallel experiments is a significant problem in amyloid disease research. Fibrils formed *in vitro* by various means may have substantially different structures; in fact many different fibril morphologies have been reported based on electron micrograph data. Due to the difficulty of fibril structural characterization, only

a handful of atomic resolution 3D structures of protein fibrils have been solved [Refs] and expedient determination of fibril form is an open problem.

Specifically in the case of α-synuclein, Guo et al. observed fibril polymorphism initially as a drift in their experimental results using *in vitro* fibrils that were propagated for several generations. The shift of results as a function of fibril generation indicated that the templated propagation was not proceeding with perfect fidelity and the spontaneous appearance of additional fibril forms or cross-seeding by the original fibril form was producing fibrils that had distinct properties in biochemical and cell-culture based assays. Their hypothesized model is shown in Figure 6.1b [Guo 2013]. To clarify the nature of this transition, they developed assays that can distinguish between fibril generations and were able to isolate two strains which seed fibrils that maintain the parent fibril's properties.

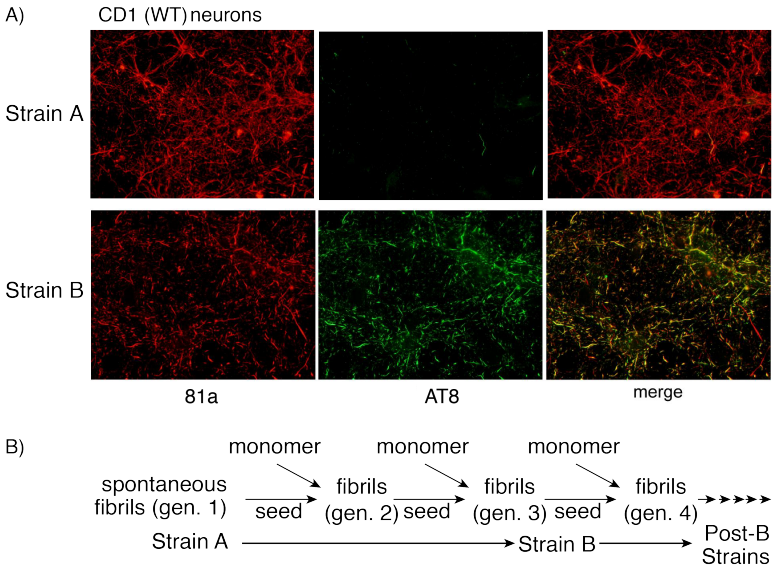

Figure 6.1 - Strain reproduction. A) CD1 neuronal cell. First row, treated with 2 μg U-^{13}C,^{15}N α-synuclein fibrils seeded with strain A fibrils. Second row treated with 2 μg U-^{13}C,^{15}N α-synuclein fibrils seeded with strain B. First column stained with 81A antibody for a-synuclein. Second column stained with AT8 antibody sensitive to tau. Third column demonstrating colocalization of tau and a-synuclein fibrils in strain B sample. The weak cross-seeding of tau fibrils by the strain A-treated sample and the strong cross- by the strain B-treated sample indicates successful propagation of strains. B) schematic indicating procedure for propagation of strains and eventual conversion of strain A into strain B.

In their work, Guo et al. observed substantial differences of the ability of the two fibril forms to cross-seed the aggregation of tau, another protein known to form fibrils and associated with neurodegenerative disease. Upon further investigation, they determined that the differences were not due to proteolysis or

chemical modification, but were more likely due to structural differences that caused the differential digestion of the fibrils by proteinase K.

In collaboration with the Lee lab, we acquired seed material of the two strains, A and B, for the purpose of performing solid-state NMR structural characterization to gain some insight into the cause of these distinct properties. This chapter describes solid-state NMR experiments conducted on $^{13}C,^{15}N$-isotopically labeled samples prepared from seeds of strain A and B α-syn pre-formed fibrils.

Materials and Methods

Pre-formed fibril (pff) samples of strain A and strain B α-syn (Guo, 2013) were provided by the Lee laboratory (Dustin Covell) to the Rienstra laboratory (Joseph Courtney). Monomeric uniformly-$^{13}C,^{15}N$-labeled (U-$^{13}C,^{15}N$) α-syn was produced at 82 mg scale and labeled fibril samples prepared for NMR by seeding with each pff strain (Kloepper, 2006). U-$^{13}C,^{15}N$ α-syn monomer was concentrated to 15 mg/mL in Dulbecco's phosphate buffered saline with 0.01% sodium azide and split between 10 1.6 mL ultracentrifuge tubes. Each tube was seeded with 0.4 mg of one pff strain (5% by mass pff seed) and vortexed to mix. Samples were incubated at 37 ºC with 200 rpm shaking for 3 weeks. All solutions gelled within 24 hours; additional incubation increases final yield of labeled fibrils. The strain A gel was cloudy but homogenous. Strain B exhibited visible striations or layers periodically below the gel surface. An aliquot of the gel material was extracted from each sample and sent back to the Lee lab for analysis where they performed proteinase K digestion and neuronal cell assays (reproducing data such as shown in Figure 6.1) indicating that the seeding was successful. The resultant fibril mass was pelleted by ultracentrifugation (130,000 g for 60 min at 4ºC), washed with a total of 2.2 mL of deionized water

in two iterations, dried under nitrogen, packed into SSNMR rotors (32 uL) and rehydrated to ~40% deionized water by mass.

By ^{13}C direct polarization measurements, the approximate quantity of ^{13}C labeled material in each sample was determined to be 20.6 mg strain A and 16.7 mg strain B.

Results and Discussion

In 1D ^{1}H-^{13}C cross polarization spectra, Strain A yields spectra of high sensitivity (Fig. 6.2a) and 2D ^{13}C-^{13}C spectra exhibit good resolution (Fig. 6.3a), consistent with a single predominant conformation and a stable, rigid fibril core. Strain B yields spectra with significantly lower sensitivity (Fig. 6.2b). In fact, while the strain B sample had 80% of the material that the strain A sample had it exhibited only 1/3 of the sensitivity. and broader linewidths but chemical shifts remarkably similar to those of strain A (Fig. 6.3b). This similarity in ^{13}C-^{13}C peak positions is especially significant given that ^{13}C chemical shifts are extremely good reporters on both secondary structure, for the Cα, Cβ, and C' chemical shifts, and environmental details such as solvent accessibility and van der Waals contacts in the core of the protein. The simplest and most likely explanation for the high similarity of the spectra is that the secondary and tertiary structures of the two fibrils are largely the same, a conclusion that is in direct opposition to the observation of different dynamics and tau binding properties.

In an effort to further characterize the structures of strains A and B, we proceeded to ^{15}N-^{13}C-^{13}C experiments for the purpose of chemical shift assignment. However, in these spectra, the differences in sensitivity were even more significant, making it infeasible to proceed with *de novo* chemical shift assignments of strain B.

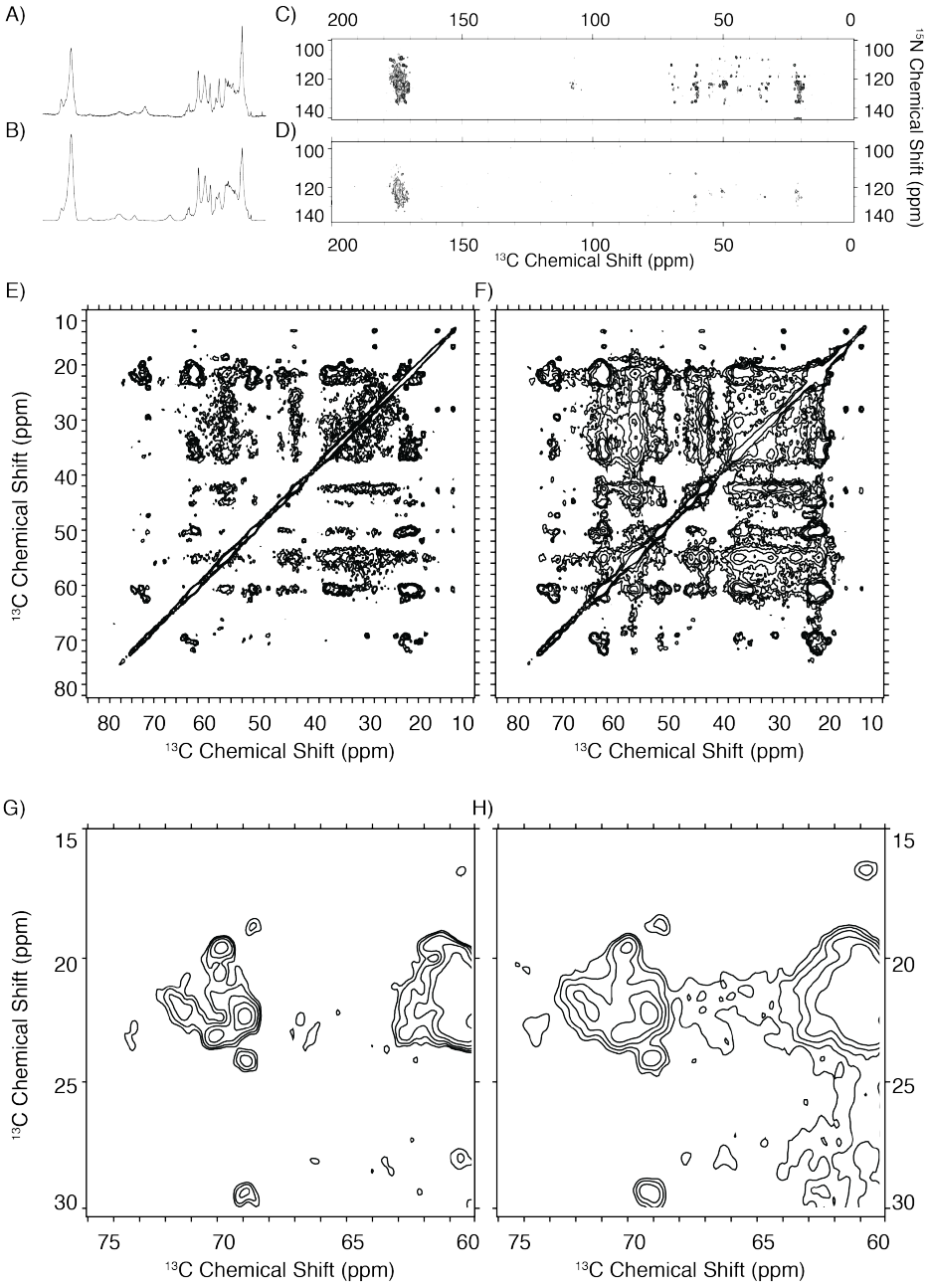

Figure 6.2 – Sensitivity and resolution comparison of strains A and B. A-B) Direct Polarization 1D ^{13}C spectra of A) Strain A and B) Strain B., C-D) 2D ^{15}N(^{13}C') ^{13}CX correlation spectra C) strain A exhibits interresidue correlations for rigid core residues. D) Strain B shows drastically lower signal intensities per unit time indicating very poor cross polarization efficiency, most likely due to increased dynamics. Spectra collected at 750 MHz proton frequency with 50 ms of ^{13}C-^{13}C DARR mixing, 80 kHz SPINAL decoupling and acquired to 7.5 ms in the ^{15}N dimension and 20 ms in the direct ^{13}C dimension. E-H) 100 ms DARR 13C-13C 2D spectra of strain A (E,G) and strain B (F,H) showing the difference in linewidths observed. E-F) full aliphatic region. G-H) threonine Cβ-Cγ region highlighting difference in linewidths.

The substantial difference in cross polarization efficiency indicates that either the ^{1}H or ^{13}C $T_1\rho$ of strain B are significantly lower because of increased dynamics at the microsecond to millisecond timescale, and linewidths between the two strains indicates that there is much more heterogeneity, either static or dynamic, in strain B than strain A. The increased mobility of strain B is consistent with it being more susceptible to dissociation of oligomers from the fibril, species indicated in the mechanism of fibril toxicity and a possible pathway for interactions with other proteins such as tau.

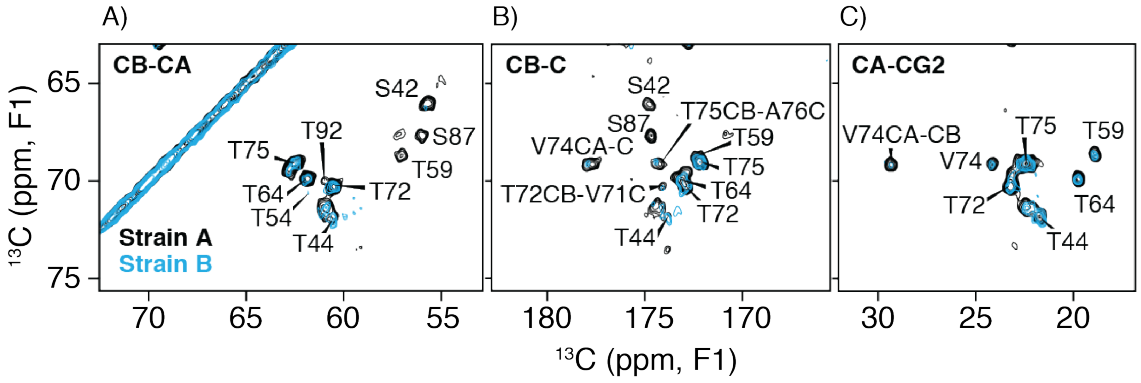

Figure 6.3 – ^{13}C-^{13}C correlation spectra of strains A and B showing intraresidue correlations. Note the excellent correspondence between peaks in strain A and strain B. A) The threonine Cβ-Ca region B) The threonine Cβ-Cγ region C) The threonine Ca-Cγ region.

Given the much lower signal intensity in strain B, we proceeded with a detailed structural study of strain A. We collected a suite of high sensitivity 2D and 3D ^{13}C-^{13}C, ^{15}N-^{13}C-^{13}C, and ^{13}C-^{15}N-^{13}C correlation spectra and with these we completed a full backbone walk and assigned the majority of the ^{13}C and ^{15}N signals in the core of the fibril (Fig. 6.4a). Upon comparison to the assignments of the Tuttle form [ref], it is apparent that the vast majority of structured residues are in different conformations, as indicated by the difference in ^{13}C chemical shifts of greater than 0.2 ppm for the majority of the core (Fig. 6.5a). Due to a well-established empirical relationship between secondary structure and chemical shifts, using those assignments we determined the secondary structure using TALOS-N (Shen, 2013) (Fig. 6.4b). The secondary structure shows the

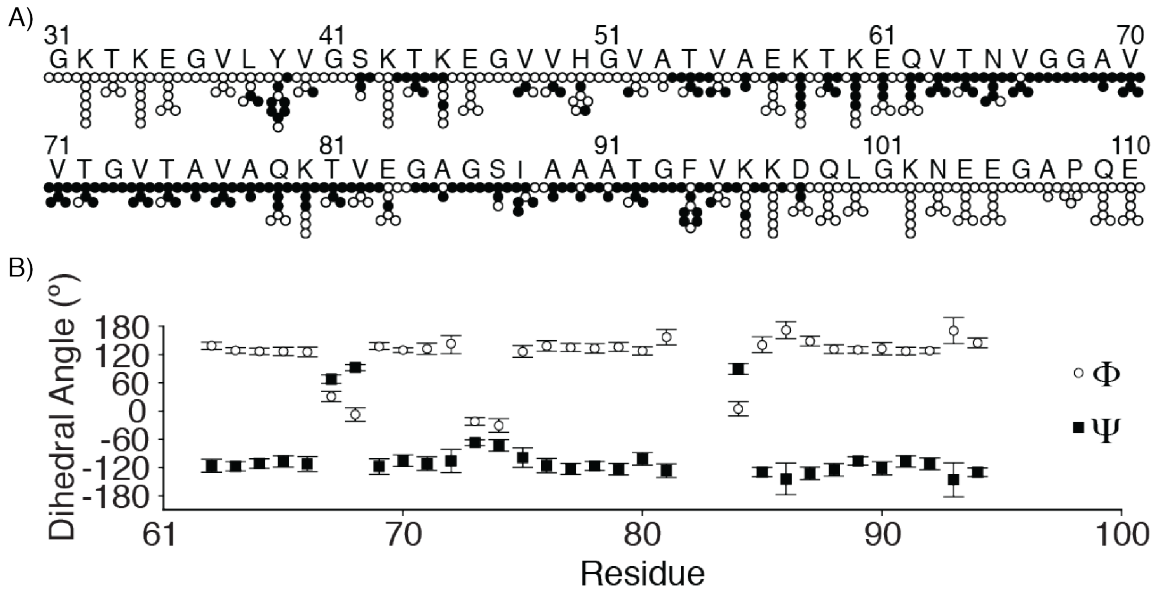

Figure 6.4 – A) Chemical shift assignment extent for uniformly ^{13}C-,^{15}N- labeled strain A. Circles represent all ^{13}C and ^{15}N atoms in each residue. Black circles have been assigned and white circles have not. B) TALOS-N predicted backbone dihedral angles showing four distinct sheets.

expected series of beta sheets but displays considerable differences from the Tuttle form (Fig. 6.5b-d). Most notably, V74, which is in the center of a core beta sheet in form K, exhibits a very unusual set of chemical shifts indicating a less often observed conformation, most likely in a β-turn. Additionally, the turn involving residues G84-A85-G86 in form K is part of a β-sheet in strain A. Overall, it is clear that the arrangement of β-sheet structures in the core differs somewhat in strain A from form K. However, there is another observation in the literature that bears some similarity to our observations of strain A. Gath et al. report a fibril form with similar breaks in beta sheet placement around the low 70s and low 80s indicating that those regions of the two fibrils may share similar structures.

While uniform ^{13}C labeling facilitates complete chemical shift assignments, sparse labeling patterns such as those resulting from the use of 2- or 1,3-^{13}C glycerol as the sole carbon source during expression afford us much better resolution due to the elimination couplings between

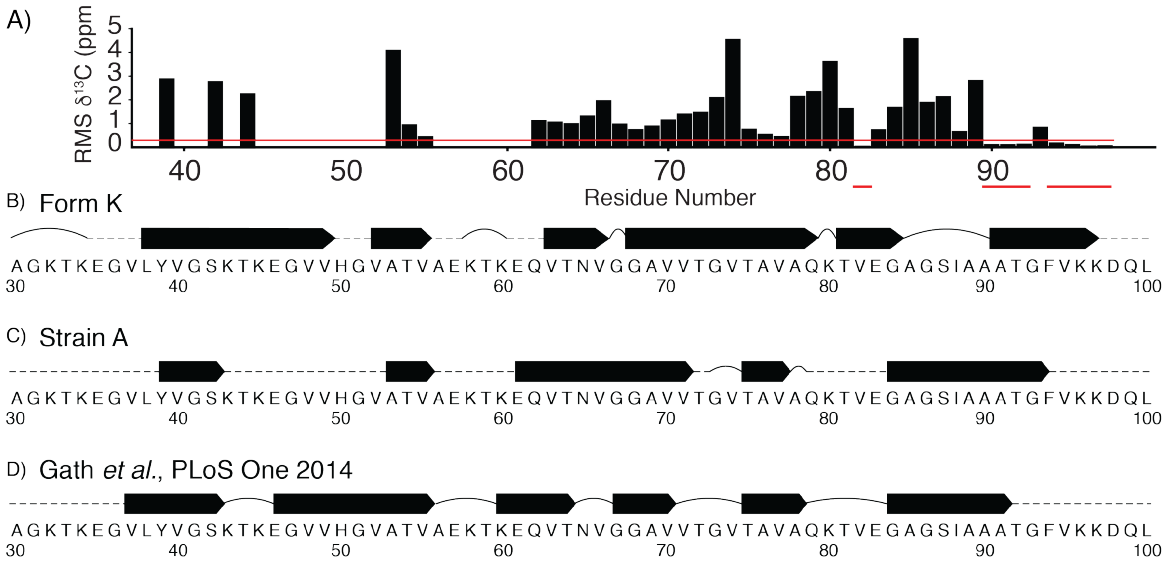

Figure 6.5 – A) Chemical shift differences between the Tuttle et al fibril form and strain A. Any difference greater than 0.3 ppm (indicated by the red line) indicate significantly different conformations. B-D) Secondary structure strip plots for different α-synuclein forms where arrows indicate beta strands, arcs represent loops, and dotted lines indicate incomplete information. B) Form K C) Strain A D) The α-synuclein fibril reported in Gath, 2014.

directly bonded carbons. Additionally, the dilution of the spin bath decreases multi-spin interactions allowing longer-range cross-peaks in correlation spectra collected for structural restraints. To ensure consistency in fibril form across samples, material from the original U-^{13}C,^{15}N labelled sample was used to seed the formation of fibrils in 2- and 1,3-^{13}C glycerol a-synuclein.

The 2-^{13}C glycerol, uniform ^{15}N (abbreviated as 2-gly) and 1,3-^{13}C glycerol uniform ^{15}N (abbreviated as 1,3-gly) expressions were performed in parallel, resulting in 34 mg of 2-gly

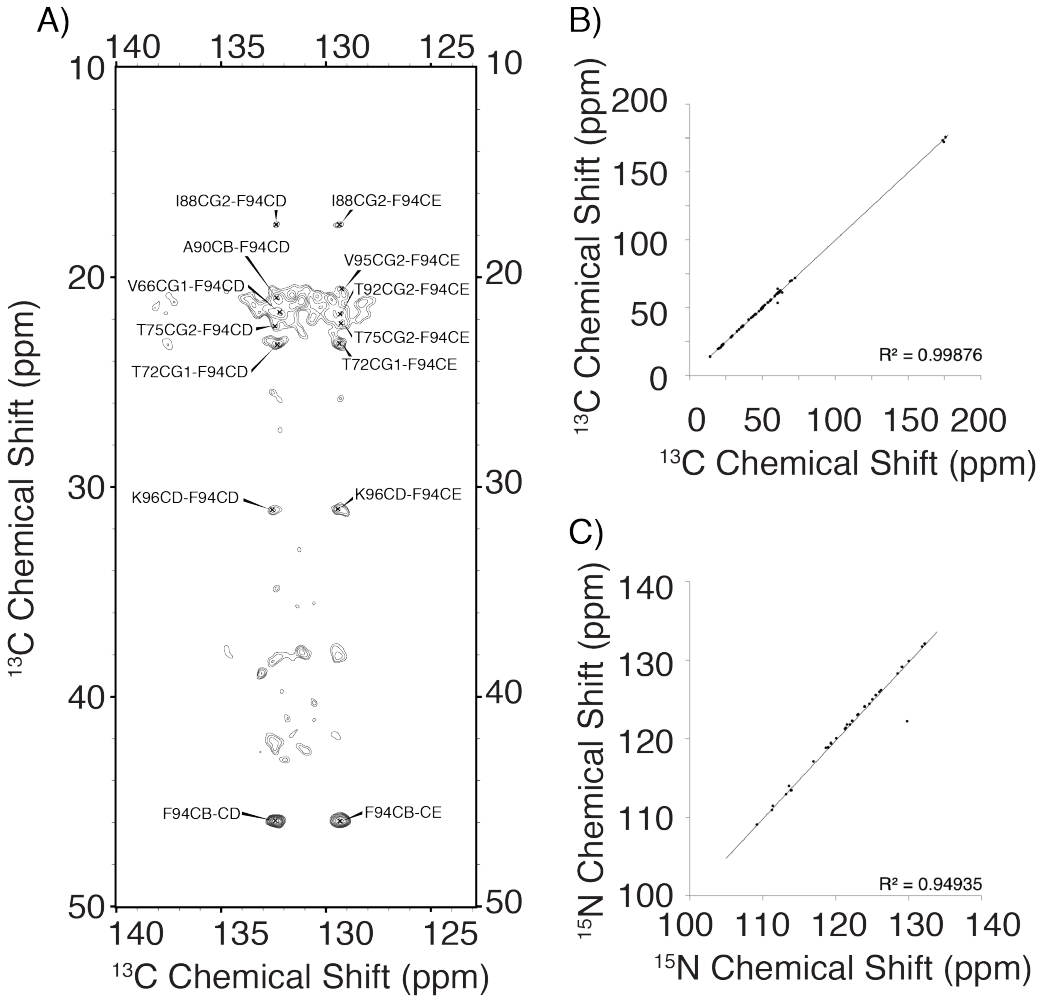

Figure 6.6 - A) Aliphatic-Aromatic correlations in 300 ms DARR spectrum of 2-gly strain A showing correlations from I88, and A90 to F94 indicating a similar hydrophobic core to the Tuttle et al. fibril form. B-C) Correspondence between UCN chemical shifts (horizontal axis) and 2-gly (vertical axis). B) ^{13}C C) ^{15}N.

monomer and 98 mg 1,3-gly monomer. Fibril samples were prepared following the same protocol as was used for the uniformly labeled samples. The fibrils were harvested as above, packed into 32 μL rotors, and brought to 40% hydration by the addition of deionized water.

We collected the standard assignment suite of SSNMR spectra on the two glycerol samples and observed that both had good sensitivity and resolution. For example, the 2-gly strain A sample showed no apparent heterogeneity and exhibited narrow lines (Fig. 6.8). Based on *de novo*

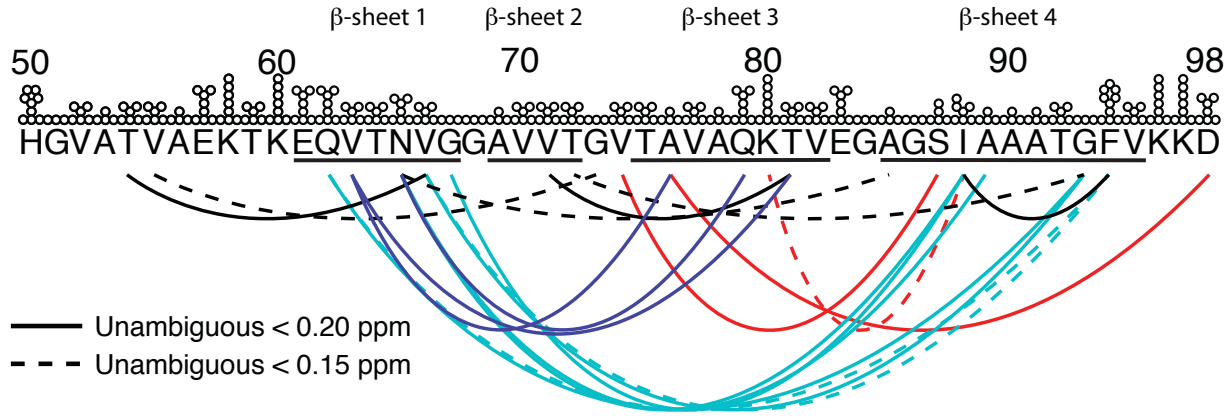

Figure 6.7 – Summary of Long-range contacts observed in strain A using UCN, 2-gly, and 1,3-gly samples. Beta sheets are indicated by underlines.

assignments of the chemical shifts in the glycerol labelled samples, we determined that they exhibit the same structure as the U-^{13}C,^{15}N sample, as demonstrated by the good agreement between the ^{13}C- and ^{15}N- chemical shifts between the U-^{13}C,^{15}N and 2-gly samples (Fig. 6.4c-d). Similar agreement is present between the U-^{13}C,^{15}N and 1,3-gly samples.

Having established that the glycerol samples seeded appropriately, we proceeded to collect long-mixing ^{13}C-^{13}C 2-dimensional and long-mixing ^{15}N-^{13}C-^{13}C spectra for the measurement of peaks corresponding to long-range through-space distances. Though, at this stage, the analysis of these spectra is incomplete, many long-range contacts have been identified. Intriguingly, the same pattern of contacts that were observed in Tuttle *et al.* indicating a hydrophobic core consisting of I88, A91, and F94 exists in strain A (Fig. 6.6). However, the pattern of other long-range contacts (Fig. 6.7) in combination with the different secondary structure indicates a substantially different fold.

Initial structure calculations indicate that the major motifs observed (LIST MOTIFS) are at least compatible and could exist in a single monomer-width fibril, as in Tuttle *et al.* (Fig. 6.8) but the pattern does not exclude the possibility of multiple monomer width fibrils.

In an attempt to reconcile the similarity of the chemical shifts of strains A and B with the differences in dynamics and pathological data, we have developed a postulated mechanistic model. We proposed that the two fibrils have nearly identical secondary and tertiary structure but that a difference in monomer stagger (as observed in amyloid beta by Tycko) or domain swapping (as observed in β-2 microglobulin, Liu, 2011) would not affect local structure significantly outside of flexible loops but could drastically change stability.

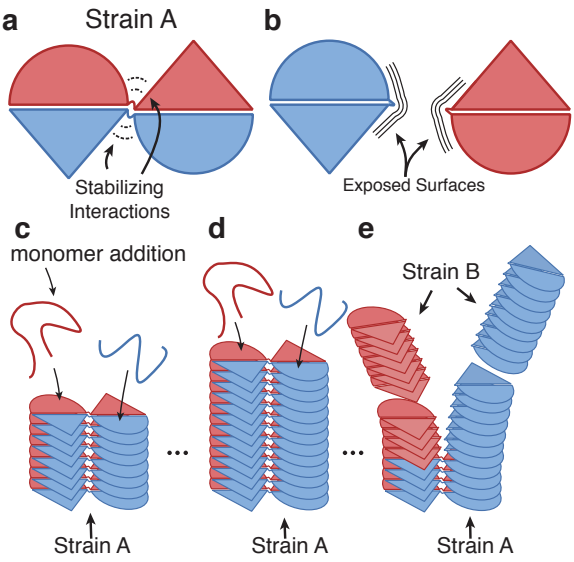

Figure 6.8 – Proposed mechanism for cross seeding of strain B by domain-swapped strain A structure.

From a thermodynamic perspective, this would seem to contradict the observation that the essentially irreversible strain A cross seeds strain B. However, if strain B is kinetically favored, the amplification process of seeded fibrillization could allow strain B to become the dominant strain while interconversion from strain B back to strain A would not occur rapidly enough build up at an appreciable quantity.

The domain-swapping hypothesis could explain the kinetically favored cross seeding from A to B if strain A contains multiple monomers per layer and B does not. As depicted in Figure 6.8, strain B could be seeded from one portion of the end face of a strain A fibril but not adopt its domain swap morphology (Fig. 6.8e). The resulting strain B fibril would retain almost identical secondary and tertiary structure to strain A but would not have potentially stabilizing inter-monomer interactions (Fig. 6.8a-b) explaining the increased mobility. Notably, the reverse cross-seeding from strain B to strain A would be exceedingly unlikely as the ends of two independent

strain B fibrils would have to exist side-by-side, in the correct orientations, long enough for enough monomer to add to the end in a domain-swapped fashion to initiate a strain A type fibril.

Conclusion and Outlook

We have thus far characterized the structural similarity and dynamic difference of strains A and B and have proposed a possible structural and mechanistic explanation for the differences observed in proteinase K digestion and neuronal cell culture. However, to test our domain-swap hypothesis and complete a full 3D structure of strain A, we will need to perform experiments on additional samples to obtain unique sets of long-range distance restraints to constrain the structure.

We plan to produce electron-microscopy samples of strain A and strain B to measure fibril width and mass-per-length, two measurements that would indicate the number of monomers for layer. Additional SSNMR samples with partial deuteration and partial labelling of methyl carbons and protons will increase resolution, decrease spectral crowding, and allow for ^{1}H-detected experiments that give us access to an additional set of long-range distances that we expect will lead to a converged atomic-resolution structure.

References

- Guo, J. L.; Covell, D. J.; Daniels, J. P.; Iba, M.; Stieber, A.; Zhang, B.; Riddle, D. M.; Kwong, L. K.; Xu, Y.; Trojanowski, J. Q.; Lee, V. M. Y., Cell, 2013, 154(1) 103-17.
- Kloepper, K. D.; Woods, W. S.; Winter, K. A.; George, J. M.; Rienstra, C. M., Protein Expr. Purif. 2006, 48, 112-117.
- Shen, Y.; Bax, A., J. Biomol. NMR. 2013, 56, 227-241.
- Saborio, G. P.; Permanne, B.; Soto, C., Nature. 2001, 411, 810-813.

- Cong Liu, Michael R Sawaya, and David Eisenberg. Nat. Struct. Mol. Biol. 2011, 18, 49–55.
- Comellas, G.; Lemkau L.R.; Nieuwkoop A. J.; Kloepper K.D.; Ladror D.T.; Ebisu, R.; Woods, W.; Lipton, A.S.; George, J. M.; Rienstra, C.M. J. Mol. Biol., 2011, 411, 881-895.
- Tuttle, M. D.; Comellas, G.; Nieuwkoop, A. J.; Covell, D. J.; Berthold, D. A.; Kloepper, K. D.; Courtney, J. M.; Kim, J. K.; Schwieters, C. D.; Lee, V. M. Y.; George, J. M.; Rienstra, C. M., Nat. Struct. Mol. Biol. 23, 5, 409-415. 2016.
- Gath, J.; Bousset, L.; Habenstein, B.; Melki, R.; Böckmann, A.; Meier, B. H., PloS One., 2014, 9(3):e90659.
- Castellani, F.; van Rossum, B.-J.; Diehl, A.; Schubert, M.; Rehbein, K.; Oschkinat, H., Nature, 2002, 420. 98-102.

CONCLUSIONS

Computational data analysis holds great promise for accelerating the field of NMR and avoiding the pitfalls of human bias and error. In this dissertation I have presented a handful of methods designed to enhance a researcher's ability to make sense of their data without requiring intricate modelling of their system or rely on manual analysis. While these methods lead to answers to very specific questions, two other areas could benefit greatly from extensions of the methods: exploratory data analysis and model criticism.

Exploratory data analysis is the search for interesting patterns in data through the use of data summarization and visualization techniques that allow the researcher to quickly observe features worth further investigation. In some types of data, identifying patters is easy. For example, given a set of images of numbers it is easy for a human to identify the relevant details, the digits from 0 to 9, and disregard the unimportant features like handwriting style and line width. Such a determination of important features is fundamental to any analysis but is surprisingly difficult to do in an automated fashion.

One class of techniques for identifying important features is dimensionality reduction where an algorithm finds a simplified representation of the data that retains most of its information content. Ideally the algorithm will distribute data in meaningful ways but often achieving a useful low-dimensional representation requires already knowing a useful description of the data.

In the first chapter, I showed that the COMPASS score is a robust measure of similarity between peak sets. It is reasonable to assume that it would also work on comparing experimental peak lists to each other. Many existing dimensionality reduction techniques only need a distance function between data points as input. One algorithm that is particularly popular for this purpose

is t-distributed stochastic neighbor embedding (t-SNE) [van der Maaten 2008] which optimizes a low-dimensional distribution of data points such that the pairwise distances are as close as possible to the distances in the original data. Effectively t-SNE squishes the data into a low-dimensional space, typically the 2D plane, in such a way that the inherent structure is preserved. Using the COMPASS score in combination with t-SNE to distribute approximately 100 peak lists reconstructed from entries in the BioMagResBank and highlighting the data points

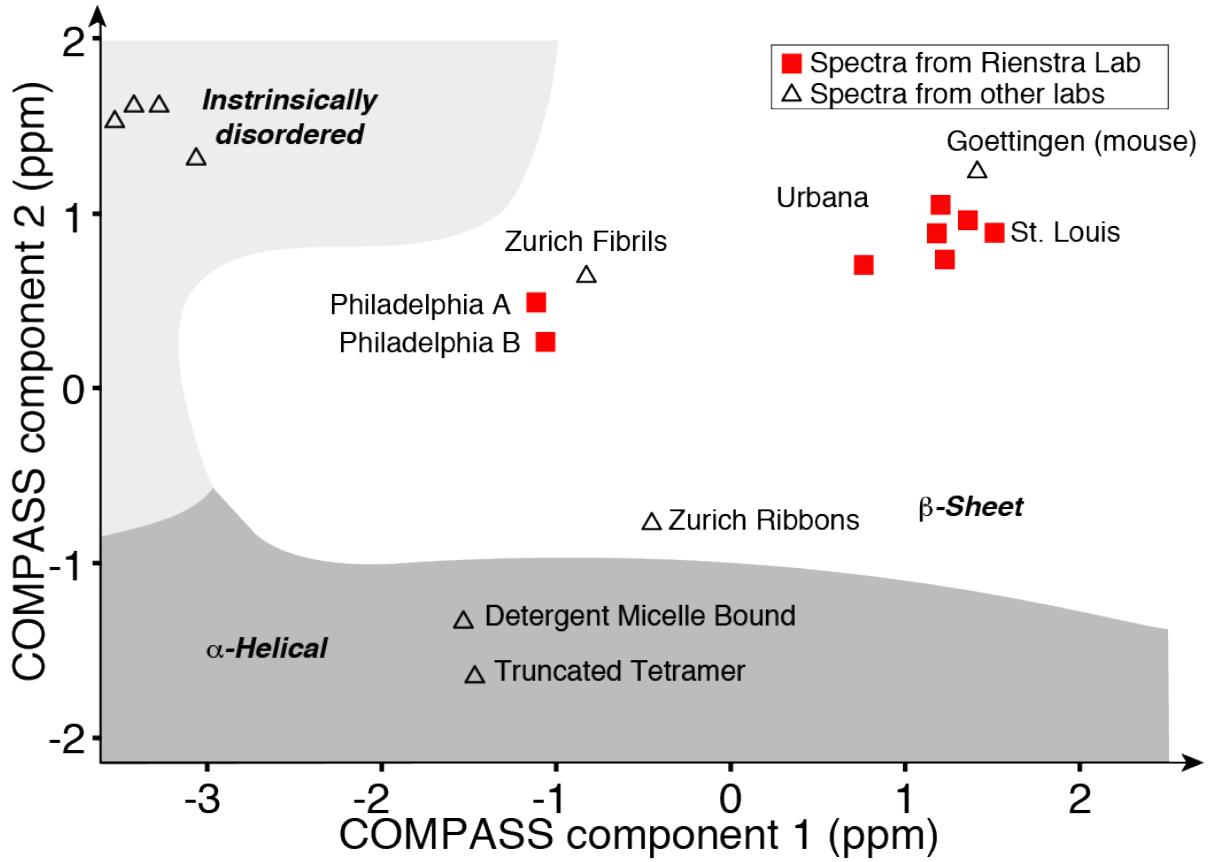

Figure C.1 – Structural Landscape.Peak lists reconstructed using the GPS method, scored for pairwise distance with the COMPASS score, and distributed in the 2D plane using t-distributed stochastic neighbor embedding reveals the natural grouping of alpha synuclein forms according to structure.

corresponding to many different forms of α-synuclein reveals a small number of distinct groupings. Intrinsically disordered and alpha helical forms were separated but so where two major groupings of fibrils. These two groupings are the same groupigns that experienced NMR spectroscopits identify after manual analysis. The ability of an objective computational method

to recover this high-level detail is somewhat surprising given the simplicity of the COMPASS score and indicates that similar exploratory data analysis may allow for the discovery of other interesting patterns in collections of NMR spectra.

While the COMPASS-t-SNE method gave interesting results it still relies on the reduction of data to a set of coordinates describing the peak positions and ignores the signal intensity and line widths that experienced researchers use extensively. Ideally, data analysis methods could operated directly on experimental data and extract similarly high-level features from the data. Based on the intuition that researchers judge spectral data primarily based on its

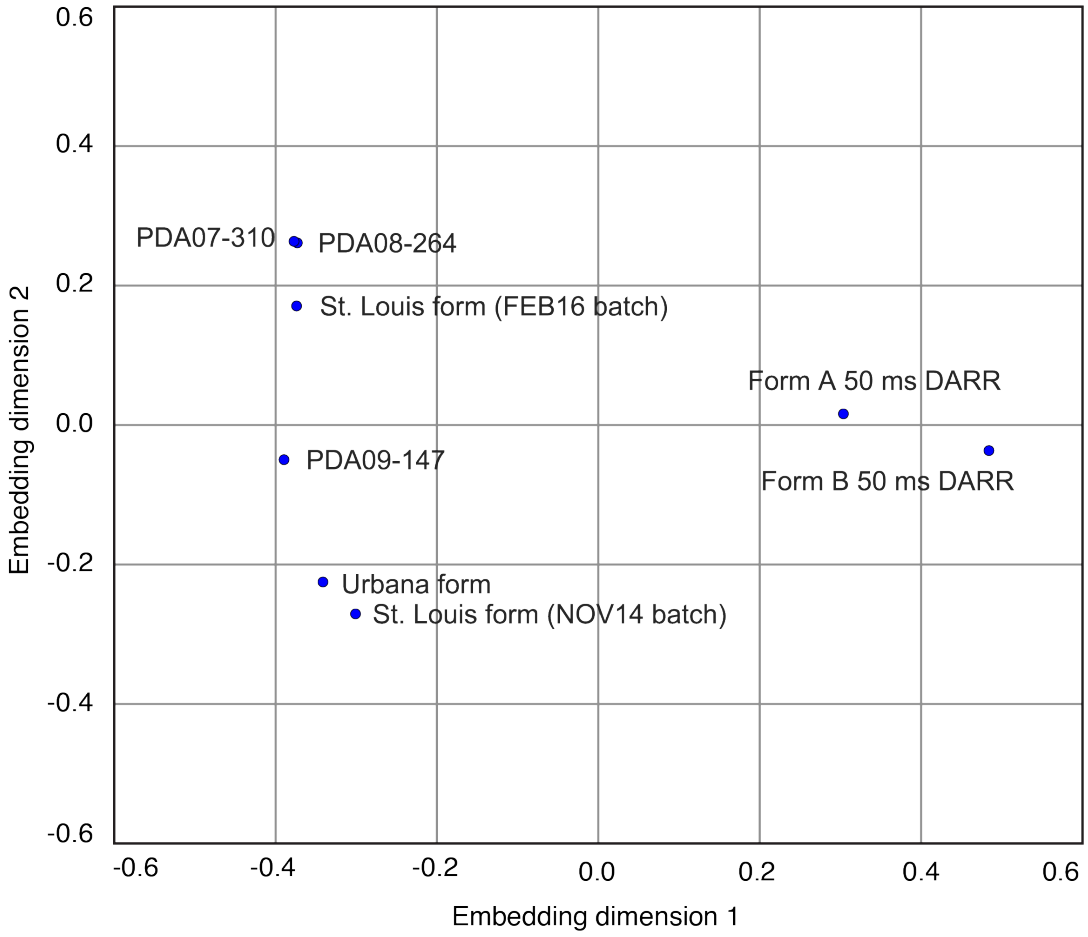

Figure C.2 – Distribution of spectra using modified locally linear embedding directly on a normalized and contrast-enhanced set of spectra. The same splitting of fibrils into two major groups is seen without picking peaks.

difference from noise I attempted to map spectral intensity to the probability that the data at a given position was signal rather than noise using a sigmoid function weighted such that data points at an intensity six times the noise are mapped to a 50% probability of being real signal. Then by removing the part of the normalized spectra that were common to an entire series of spectra I was able to produced ~1,000,000 dimensional vectors for using in dimensionality reduction techniques. Utilizing one of the earliest nonlinear dimensionality techniques, Locally Linear Embedding, I mapped the spectral vectors into the 2D plane (Fig. C.2).

The resulting distribution of spectra from the described contrast enhancement procedure recovers the grouping of fibrils into two major families, the Tuttle *et al* form including forms identified in Urbana and St. Louis, and the strain A/B family. The success of these two preliminary attempts to perform high-level analysis of NMR data by computational means indicate that this is a fertile area deserving of extensive research.