STRUCTURAL DETERMINANTS OF CARDIAC LIGHT CHAIN AMYLOIDOSIS

BY

DENNIS WILLIAM PIEHL

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Biochemistry
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Professor Chad M. Rienstra, Chair
Professor Robert B. Gennis
Professor Satish K. Nair
Professor Martin H.W. Gruebele

ABSTRACT

Light chain amyloidosis (AL) is an acute systemic amyloid disease in which the overexpression and misfolding of immunoglobulin light chains leads to the formation of insoluble fibrils at critical organs in the body. The typical prognosis for AL patients is extremely severe—usually less than three years of diagnosis—but in the case of cardiac involvement, this is reduced to under a year. As a result of the complex and diverse nature of the disease, our understanding of how it onsets and imparts its devastating effects remains elusive, consequently delaying progress on the identification of robust diagnostic and therapeutic strategies. To address this deficiency, we investigate the structural and dynamical properties of the fibrils involved in a severe case of cardiac AL disease, formed from the light-chain variable domain (V_L) patient-derived protein, AL-09. We performed these studies through the use of solid-state nuclear magnetic resonance (SSNMR) spectroscopy, which have revealed new insights on specific features of the fibril species. First, we report our findings on the ^{13}C and ^{15}N chemical shift assignments and sequence involvement of AL-09 V_L fibrils, followed by subsequent investigations on the predominant secondary structure and relative dynamics across the protein sequence. In addition, we present a comparison of AL-09 V_L fibrils (prepared *in vitro*) with *ex vivo* amyloid deposits of another AL protein fibril (HIG) obtained from human spleen tissue using 1D ^{13}C SSNMR analysis. Next, we describe the development and implementation of 4D ^{13}C-detected SSNMR experiments that combine non-uniform sampling with band-selective J-decoupling pulses to enable the resolution of previously unassignable resonances while maintaining high sensitivity. Last, we introduce an improved program and strategy for the computationally-aided assignment of protein chemical shifts that incorporates peak intensity and frequency overlap information as an additional scoring function in the multi-objective search for optimal assignment solutions.

ACKNOWLEDGEMENTS

In reflecting back on the people who have helped guide me here today and ultimately make this work possible, I am of course flooded with an overwhelming number of names that have all in some way imparted an influence on my personal and academic experiences. It goes without saying that I am incredibly grateful and fortunate to have had so many wonderful people steer me in this direction. Nevertheless, there are certainly a select handful of individuals who have made a profound impression on not only my intellectual or professional growth, but also my personal and scientific interests, perspectives, and thought process. These are the teachers, colleagues, family members, and friends to whom I express all of my thanks here.

I would like to begin by acknowledging the professors who inspired me to pursue graduate school in the first place. As an undergraduate biochemistry student at Seattle University, I had the great fortune of being placed under the advising of Prof. Jennifer Loertscher. It is because of Prof. Loertscher that I had the opportunity to not only study abroad in Germany during my junior year of college, but to pursue a research internship at the University of Leipzig the summer prior as well. That whole experience became one of the most influential and formative periods of my life, and I could never be more thankful for having the enormous support from Prof. Loertscher to make it all happen. In addition, I want to thank my undergraduate research advisor, Prof. Ian Suydam, for mentoring me through the course of my first extended research experience as well as impressing onto me a deeper understanding of the scientific method during the process. The research I performed in Prof. Suydam's laboratory played an influential role in my decision to pursue a Ph.D. in biochemistry, and I appreciate all of the support that he provided to help lead me here.

This now brings me to acknowledging the person who trained and guided me throughout the course of my graduate research work, Professor Chad Rienstra. Along with teaching me everything I know about NMR, Prof. Rienstra presented me with several particularly prominent opportunities. As an example, he showed me how to build a new project from the ground up, and carry it forward in an

efficient and productive manner by making important judgement calls at each step along the way. There were many pivotal moments during the course of my Ph.D. where I could have (and likely would have, regrettably) continued persisting on some problem for another year or more before realizing that I may be heading into a dead end. Thankfully, however, the majority of these situations were avoided because of the insightful direction that Prof. Rienstra was able to offer through our discussions of the data. Moreover, it was through these meetings in which I learned how to critically analyze the data that I have, not just for obtaining the information that it holds, but in terms of its significance and relevance to the greater focus of the project. The progress and success of my thesis work would not have been possible without this guidance.

Another set of experiences that I am especially grateful for are all of the hypothetical discussions that I was able to have with Prof. Rienstra during our meetings, as well as the rest of my doctoral committee, Prof. Bob Gennis, Prof. Satish Nair, and Prof. Martin Gruebele, at my examinations. Generally, the way these thought experiments proceeded would be with a question being posed about how to test a certain hypothesis, followed by me attempting to propose some way of addressing it, and then receiving advice back with a better way of doing it. Irrespective of whether I was right or wrong, I have always deeply enjoyed these types of discussions, for two reasons. First, each question and answer taught me something new. Second, it made me truly appreciate the fact that for almost any question you ever want to answer, there is some sort of experiment or series of tests that you can perform to get the information you need to figure it out. This type of thinking is what I consider to be one of the most valuable assets that I have gained during my graduate career, and thus am widely thankful to have had these experiences.

Alongside the support from my advising committee, I owe a tremendous amount of gratitude towards all of the collaborators who made this work possible. First and foremost, I want to thank Prof. Marina Ramirez-Alvarado and Dr. Luis Blancas-Mejía, for not only being responsible for initiating our collaboration and proposing the idea for the project in the first place, but for their vital contributions to its success by providing us with the isotopically-labeled fibril samples that were used in our studies. I am

also truly appreciative of their enthusiasm and efforts towards maintaining such a strong collaboration, both though all of our conference calls as well as hosting me at the Mayo Clinic to provide me with training. Furthermore, I would like to thank our collaborators at the University of Tennessee Medical Center, Prof. Jonathan Wall and Prof. Stephen Kennel, for providing us with the light-chain amyloid tissue sample that was used in our studies.

A significant amount of my graduate research work was also only made possible by the support of my colleagues here at the University of Illinois at Urbana-Champaign, to whom a great deal of gratitude is due. In particular, among my fellow lab members, I thank Alexander Barclay, Lisa Della Ripa, Manali Ghosh, Adedolapo Ojoawo, and Dr. Agnieszka Lewandowska, for assisting with the execution, monitoring, and analysis of many of the key experiments that were performed in our studies. I would also like to acknowledge Dr. Joseph Courtney, Dr. Xiangyan Shi, and Dr. Alexander Greenwood, for contributing to multiple aspects of the research related to experimental design and data analysis methods. Additionally, I extend much appreciation to Dr. Deborah Berthold for her training and support on sample preparation procedures. Beyond our research group, I am also very grateful for the collaborative support from Michael Hallock and Dr. Taras Pogorelov, for their valuable discussions and contributions to our research efforts.

Finally, I would like to express my overwhelming gratitude to my family for supporting and encouraging me all this way. I could not see myself in this position today if it were not for the influence of my older brother, Kevin, who in many ways served as my mentor during the first 18 years of my life. He has always inspired me to be curious about things, and continues to be someone whom I look up to greatly. I am incredibly grateful for my younger brother, James, whom I have always looked forward to being able to catch up with when I visit home, and who continues to impress me with the amazing person that he has become every time I do. I am thrilled to see where he finds himself in the future, and I could never be more confident in what he decides to pursue. Most of all, I would not be anywhere near this point it were not for the unconditional and loving support of my parents, Cathy and Mike. Not only have they continued to walk with me all along the way, supporting me at every milestone and every decision

that I made to lead me here, but they also never responded with anything less than encouragement. Even despite all the rather shortened visits back home while away during my undergraduate and graduate life, they have remained patient with me and continued to motivate me in my endeavors. Thank you for giving me the faith and support to follow the path that I wanted to pursue.

To my mother and father

TABLE OF CONTENTS

Chapter 1: Introduction 1

Chapter 2: Solid-state NMR Chemical Shift Assignments for AL-09 V_L Immunoglobulin Light Chain Fibrils 10

Chapter 3: Immunoglobulin Light Chains Form an Extensive and Highly Ordered Fibril Involving the N- and C-termini. 25

Chapter 4: Distinct Structural Features of Light-Chain Fibrils Revealed by Four-Dimensional ^{13}C-Detected Solid-State NMR Spectroscopy 55

Chapter 5: Towards a Robust and Streamlined Strategy for Assigning Chemical Shifts in Proteins Using Peak Intensity and Overlap-Weighted Scoring on NUS SSNMR Data 87

Chapter 6: Conclusions and Outlook 136

CHAPTER 1: INTRODUCTION

Amyloid diseases encompass a broad group of disorders that are characterized by the misfolding and aggregation of proteins in the body and their subsequent deposition as elongated, insoluble fibrils at various organs. The way these diseases are typically classified is based on the location at which the fibrils involved deposit in the body. In general, they can be separated into two main categories: neurodegenerative amyloid diseases, in which fibril formation occurs in the brain tissue of patients, and non-neuropathic amyloid diseases, in which fibrils deposit at other organs of the body apart from the neural tissue[1,2]. The pathology that occurs in the class of neurodegenerative diseases is presented in cases such as Parkinson's disease, Alzheimer's disease, Huntington's disease, as well as many other prevalent disorders[3]. For non-neuropathic amyloid diseases—more generally known as amyloidoses—this pathology can manifest as either localized amyloidosis, in which fibril build-up occurs specifically at the tissue where the protein is produced, or systemic amyloidosis, in which multiple organs and tissues can be targeted[4].

A significant amount of research, especially in the past few decades, has been directed at improving our understanding of the underlying mechanisms of amyloid fibril formation in these diseases, as well as how to more effectively diagnose and treat them[5]. To capture a complete picture of the determinants driving the progression of amyloid diseases, researchers have applied a wide array of techniques to study them from many different angles, ranging from genetic regulation of the precursor protein[6,7], mathematical modeling of fibril formation kinetics[8], thermodynamics of protein folding[9,10], cellular toxicity assays[11-13], and continuing onto their testing in animal models[14] and eventually clinical drug trials[15]. Each of these methods provides a different insight into the amyloid formation process. One technique that is particularly well-suited for studying both the molecular-level structure and dynamics of amyloid fibrils is magic-angle spinning (MAS) solid-state nuclear magnetic resonance (SSNMR) spectroscopy, in which there is no inherent molecular size limitation and is specifically equipped to study insoluble materials. Indeed, MAS SSNMR has been applied to the cases of many disease-related fibrils,

and has enabled the elucidation of the three-dimensional structure for multiple amyloid systems in only the past few decades, including those involved in neurodegenerative diseases such as Alzheimer's disease and Parkinson's disease[16-22], localized amyloidoses such as that often presented in type II diabetes mellitus[23,24], as well as localized or systemic dialysis-mediated amyloidosis[25,26]. The availability of these structures has provided foundational insight into the unique conformational fold that each fibril structure exhibits, the types of interactions that hold the fibril core together, as well as the intermolecular interactions that support the structural integrity of the extended fibril[27]. Furthermore, the elucidation of these structures can also provide a basis on which the design of molecules that can bind to specific sites on the fibril of interest can be investigated. Indeed, relatively recent work has already shown that this can be done using a library of compounds to computationally sample and score their binding to the β-amyloid fibril structure[28,29]. Clearly, there is tremendous value in having access to this structural information, especially when combined the details obtained from all of the other methods. As such, we now turn to the case of systemic amyloidoses, which comprise the group of amyloid diseases which has received notably less attention from structural pursuits, in which we apply and further develop several of the techniques used for the fibril systems described above.

The most common form of systemic amyloidosis is presented by the disease light chain amyloidosis (AL)[4]. AL is an acute systemic disease in humans that is described by the overexpression of immunoglobulin light-chain proteins throughout the bloodstream which—upon misfolding—form into insoluble fibrils and deposit at critical organs of the body, culminating in tissue damage and death[30]. The average prognosis for patients suffering from AL disease is generally within two to three years of diagnosis, but is reduced to under a year if the afflicted organ is the heart, which is presented by approximately 50% of all AL cases[31-33]. Due to the highly variable nature of immunoglobulin light chain sequences and wide range of possible tissue sites for fibril deposition, the search for a working animal model as well as effective diagnostic and therapeutic strategies for AL has remained largely limited[4]. These efforts have been further hindered by the absence of a full understanding of the molecular mechanisms governing the fibril formation process as well as the three-dimensional fibril structure, both

of which would inform the rational design of new treatments or diagnostic tools[34-36]. Accordingly, here we present our collection of investigations on the structural determinants of cardiac AL disease, including studies of both the fibril structure and dynamics, as pursued through MAS SSNMR spectroscopy.

The work described in the following chapters is concentrated on the light-chain variable domain (V_L) of the protein, AL-09, which represents one particularly severe case of cardiac AL disease[31,37]. AL-09 V_L is a 108 amino acid residue protein that contains seven mutations relative to the κI O18/O8 germline sequence from which it is derived[37]. Of these mutations, Y87H, N34I, and K42Q are non-conservative and are known to contribute the most to the overall increased propensity and rate of fibril formation[30,38]. To explore the involvement of these residues and other residues in the fibril structure, our initial studies involved the acquisition and analysis of a series of multidimensional SSNMR ^{13}C and ^{15}N correlation spectra on uniformly [^{13}C,^{15}N]-labeled AL-09 V_L fibrils. In addition to enabling site-specific chemical shift assignments to be made for nearly half of the backbone and side chain atoms in AL-09 V_L (via a sequential backbone walk), our data reveal that approximately 70 residues of sequence in total are observed to be structured in the fibril species[39]. Further, the secondary chemical shift trends and backbone torsion angle predictions determined from these assignment data revealed that the structured residues of AL-09 V_L fibrils assume a predominantly β-sheet conformation[39,40]. Subsequently, we proceeded to investigate the dynamics of the fibril structure level through the use of ^{1}H-^{13}C dipolar recoupling experiments (via 2D T-MREV[41,42] and 3D R-symmetry[43] pulse sequences), allowing for the measurement of the scaled ^{1}H-$^{13}C\alpha$ dipolar-coupling values at both the sequence-wide and site-specific levels. Our results indicated that AL-09 V_L fibrils exhibit high rigidity across the majority of the protein sequence, and moreover that this rigidity is particularly evident for residues near the N- and C-termini[40]. Together, these data suggest that the N- and C-terminal regions are primarily involved in maintaining the rigid fibril structure, while the more central regions of the sequence—i.e., where most of the mutation sites occur—may play a larger role in destabilizing the native protein structure to allow the initial misfolding event and fibril formation to occur.

To assess whether the structural features we observe for AL-09 V_L fibrils (prepared *in vitro*) are common to other AL fibrils, we compared the ^{13}C 1D spectrum to that acquired for an *ex vivo* AL amyloid sample (obtained from post-mortem human spleen tissue; HIG). Interestingly, the samples demonstrate qualitatively similar spectral features and secondary chemical shift trends, while the presence of several unique peaks in HIG spectrum are indicative of the presence (and likely association) of tightly bound glycosaminoglycans from the extracellular matrix of the spleen tissue[40]. These results will serve as a framework on which the structural study of isotopically-enriched and seeded HIG fibrils can be performed.

In parallel with our above studies, we also describe the design and application of two new methods for SSNMR data collection and analysis. The first is the implementation of ^{13}C-detected 4D experiments that provide internuclear correlation information between the *i* and *i*-1 residues (via $^{13}C\alpha_i$-$^{15}N_i$-$^{13}C'_{i-1}$-$^{13}CX_{i-1}$ transfer) or *i* and *i*+1 residues (via $^{13}C'_{i-1}$-$^{15}N_i$-$^{13}C\alpha_i$-$^{13}C'_i$ transfer)[44], combined with band-selective J-decoupling pulses on the indirect ^{13}C-dimensions to deliver further resolution enhancement without compromising sensitivity[45]. Furthermore, we have implemented these experiments with the use of non-uniform sampling (NUS), a technique in which only a subset of the hypercomplex data points are collected and—when combined with powerful data reconstruction tools such as SMILE[46]—allow for efficient acquisition of high-dimensionality spectra with enhancements to both resolution and sensitivity[47,48]. Following initial performance testing of the pulse sequence with GB1, we proceed to apply this to the case of AL-09 V_L fibrils and demonstrate the ability to resolve and assign chemical shift resonances that were previously inaccessible at lower dimensionality, in addition to reducing the total experiment time by a factor of 50 or more. Moreover, our results on AL-09 V_L also reveal that several residues the central region of the protein sequence are also partially structured in the fibril form and exhibit a unique set of chemical shifts, suggesting that they may adopt a distinct conformation relative to the rest of the AL-09 V_L fibril structure. This work comprises the primary focus of Chapter 4.

The second method we introduce (which forms the topic of Chapter 5), is an improved program for computationally-assisting the assignment of chemical shifts in protein SSNMR data. Specifically, we have incorporated peak intensity and frequency overlap information as a new objective within the Monte Carlo and multi-objective approach to identifying optimal assignment solutions which have been reported previously[49-51]. Through this, we essentially provide a method of scaling the overall quality of a given assignment solution based on an additional aspect of experimental data, which we demonstrate offers a significant improvement to the robustness of the program relative to the previous programs. In particular, when provided with data containing many inter-residue cross-peak correlations within the same spin systems (e.g., a list of frequencies representing a single residue, plus several non-representative frequencies), our program consistently performs with a much higher probability of success than previous programs. Furthermore, we present several additional computational tools that streamline the preparation of input to the new assignment program directly from the data processing step, without requiring time-consuming manual analysis of the acquired spectrum. In addition to applying this strategy to GB1 with nearly perfect assignment success, we apply it to the challenging case of AL-09 V_L fibril data and show that the program is able to accurately assign the majority of the residues which have previously been assigned manually with moderate consistency, even with minimally analyzed data sets.

1.1. REFERENCES

(1) Chiti, F.; Dobson, C. M. Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.* **2006**, *75*, 333–366.

(2) Chiti, F.; Dobson, C. M. Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade. *Annu. Rev. Biochem.* **2017**, *86*, 27–68.

(3) Hartl, F. U. Protein Misfolding Diseases. *Annu. Rev. Biochem.* **2017**, *86*, 21–26.

(4) Blancas-Mejía, L. M.; Ramirez-Alvarado, M. Systemic amyloidoses. *Annu. Rev. Biochem.* **2013**, *82*, 745–774.

(5) Comellas, G.; Rienstra, C. M. Protein structure determination by magic-angle spinning solid-state NMR, and insights into the formation, structure, and stability of amyloid fibrils. *Annu. Rev. Biophys.* **2013**, *42*, 515–536.

(6) Arendt, B. K.; Ramirez-Alvarado, M.; Sikkink, L. A.; Keats, J. J.; Ahmann, G. J.; Dispenzieri, A.; Fonseca, R.; Ketterling, R. P.; Knudson, R. A.; Mulvihill, E. M.; Tschumper, R. C.; Wu, X.; Zeldenrust, S. R.; Jelinek, D. F. Biologic and genetic characterization of the novel amyloidogenic lambda light chain-secreting human cell lines, ALMC-1 and ALMC-2. *Blood* **2008**, *112*, 1931–1941.

(7) Niblock, M.; Gallo, J. M. Tau alternative splicing in familial and sporadic tauopathies. *Biochem. Soc. Trans.* **2012**, *40*, 677–680.

(8) Knowles, T. P.; Waudby, C. A.; Devlin, G. L.; Cohen, S. I.; Aguzzi, A.; Vendruscolo, M.; Terentjev, E. M.; Welland, M. E.; Dobson, C. M. An analytical solution to the kinetics of breakable filament assembly. *Science* **2009**, *326*, 1533–1537.

(9) Kachlishvili, K.; Dave, K.; Gruebele, M.; Scheraga, H. A.; Maisuradze, G. G. Eliminating a Protein Folding Intermediate by Tuning a Local Hydrophobic Contact. *J. Phys. Chem. B* **2017**, *121*, 3276–3284.

(10) Bhattacharyya, A. M.; Thakur, A. K.; Wetzel, R. Polyglutamine aggregation nucleation: thermodynamics of a highly unfavorable protein folding reaction. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 15400–15405.

(11) Luk, K. C.; Song, C.; O'Brien, P.; Stieber, A.; Branch, J. R.; Brunden, K. R.; Trojanowski, J. Q.; Lee, V. M. Exogenous α-synuclein fibrils seed the formation of Lewy body-like intracellular inclusions in cultured cells. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 20051–20056.

(12) Volpicelli-Daley, L. A.; Luk, K. C.; Patel, T. P.; Tanik, S. A.; Riddle, D. M.; Stieber, A.; Meaney, D. F.; Trojanowski, J. Q.; Lee, V. M. Exogenous α-synuclein fibrils induce Lewy body pathology leading to synaptic dysfunction and neuron death. *Neuron* **2011**, *72*, 57–71.

(13) Desplats, P.; Lee, H. J.; Bae, E. J.; Patrick, C.; Rockenstein, E.; Crews, L.; Spencer, B.; Masliah, E.; Lee, S. J. Inclusion formation and neuronal cell death through neuron-to-neuron transmission of α-synuclein. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 13010–13015.

(14) Luk, K. C.; Kehm, V.; Carroll, J.; Zhang, B.; O'Brien, P.; Trojanowski, J. Q.; Lee, V. M. Pathological α-synuclein transmission initiates Parkinson-like neurodegeneration in nontransgenic mice. *Science* **2012**, *338*, 949–953.

(15) Graham, W. V.; Bonito-Oliva, A.; Sakmar, T. P. Update on Alzheimer's Disease Therapy and Prevention Strategies. *Annu. Rev. Med.* **2017**, *68*, 413–430.

(16) Paravastu, A. K.; Leapman, R. D.; Yau, W. M.; Tycko, R. Molecular structural basis for polymorphism in Alzheimer's β-amyloid fibrils. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 18349–18354.

(17) Qiang, W.; Yau, W. M.; Luo, Y.; Mattson, M. P.; Tycko, R. Antiparallel beta-sheet architecture in Iowa-mutant β-amyloid fibrils. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 4443–4448.

(18) Lu, J. X.; Qiang, W.; Yau, W. M.; Schwieters, C. D.; Meredith, S. C.; Tycko, R. Molecular structure of β-amyloid fibrils in Alzheimer's disease brain tissue. *Cell* **2013**, *154*, 1257–1268.

(19) Schutz, A. K.; Vagt, T.; Huber, M.; Ovchinnikova, O. Y.; Cadalbert, R.; Wall, J.; Guntert, P.; Bockmann, A.; Glockshuber, R.; Meier, B. H. Atomic-resolution three-dimensional structure of amyloid beta fibrils bearing the Osaka mutation. *Angew. Chem. Int. Edit. Engl.* **2015**, *54*, 331–335.

(20) Colvin, M. T.; Silvers, R.; Ni, Q. Z.; Can, T. V.; Sergeyev, I.; Rosay, M.; Donovan, K. J.; Michael, B.; Wall, J.; Linse, S.; Griffin, R. G. Atomic Resolution Structure of Monomorphic Aβ42 Amyloid Fibrils. *J. Am. Chem. Soc.* **2016**, *138*, 9663–9674.

(21) Qiang, W.; Yau, W. M.; Lu, J. X.; Collinge, J.; Tycko, R. Structural variation in amyloid-beta fibrils from Alzheimer's disease clinical subtypes. *Nature* **2017**, *541*, 217–221.

(22) Tuttle, M. D.; Comellas, G.; Nieuwkoop, A. J.; Covell, D. J.; Berthold, D. A.; Kloepper, K. D.; Courtney, J. M.; Kim, J. K.; Barclay, A. M.; Kendall, A.; Wan, W.; Stubbs, G.; Schwieters, C. D.; Lee, V. M. Y.; George, J. M.; Rienstra, C. M. Solid-state NMR structure of a pathogenic fibril of full-length human α-synuclein. *Nat. Struct. Mol. Biol.* **2016**, *23*, 409–415.

(23) Nielsen, J. T.; Bjerring, M.; Jeppesen, M. D.; Pedersen, R. O.; Pedersen, J. M.; Hein, K. L.; Vosegaard, T.; Skrydstrup, T.; Otzen, D. E.; Nielsen, N. C. Unique identification of supramolecular structures in amyloid fibrils by solid-state NMR spectroscopy. *Angew. Chem. Int. Edit. Engl.* **2009**, *48*, 2118–2121.

(24) Westermark, P.; Wernstedt, C.; Wilander, E.; Hayden, D. W.; O'Brien, T. D.; Johnson, K. H. Amyloid fibrils in human insulinoma and islets of Langerhans of the diabetic cat are derived from a neuropeptide-like protein also present in normal islet cells. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 3881–3885.

(25) Iwata, K.; Fujiwara, T.; Matsuki, Y.; Akutsu, H.; Takahashi, S.; Naiki, H.; Goto, Y. 3D structure of amyloid protofilaments of β2-microglobulin fragment probed by solid-state NMR. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 18119–18124.

(26) Dogan, A. Amyloidosis: Insights from Proteomics. *Annu. Rev. Pathol.* **2017**, *12*, 277–304.

(27) Eisenberg, D. S.; Sawaya, M. R. Structural Studies of Amyloid Proteins at the Molecular Level. *Annu. Rev. Biochem.* **2017**, *86*, 69–95.

(28) Jiang, L.; Liu, C.; Leibly, D.; Landau, M.; Zhao, M. L.; Hughes, M. P.; Eisenberg, D. S. Structure-based discovery of fiber-binding compounds that reduce the cytotoxicity of amyloid beta. *eLife* **2013**, *2*, e00857.

(29) Sievers, S. A.; Karanicolas, J.; Chang, H. W.; Zhao, A.; Jiang, L.; Zirafi, O.; Stevens, J. T.; Munch, J.; Baker, D.; Eisenberg, D. Structure-based design of non-natural amino-acid inhibitors of amyloid fibril formation. *Nature* **2011**, *475*, 96–103.

(30) Baden, E. M.; Sikkink, L. A.; Ramirez-Alvarado, M. Light chain amyloidosis – current findings and future prospects. *Curr. Protein Pept. Sci.* **2009**, *10*, 500–508.

(31) Abraham, R. S.; Geyer, S. M.; Price-Troska, T. L.; Allmer, C.; Kyle, R. A.; Gertz, M. A.; Fonseca, R. Immunoglobulin light chain variable (V) region genes influence clinical presentation and outcome in light chain-associated amyloidosis (AL). *Blood* **2003**, *101*, 3801–3808.

(32) Levinson, R. T.; Olatoye, O. O.; Randles, E. G.; Howell, K. G.; DiCostanzo, A. C.; Ramirez-Alvarado, M. Role of mutations in the cellular internalization of amyloidogenic light chains into cardiomyocytes. *Sci. Rep.* **2013**, *3*, 1–8.

(33) McLaughlin, R. W.; De Stigter, J. K.; Sikkink, L. A.; Baden, E. M.; Ramirez-Alvarado, M. The effects of sodium sulfate, glycosaminoglycans, and Congo red on the structure, stability, and amyloid formation of an immunoglobulin light-chain protein. *Protein Sci.* **2006**, *15*, 1710–1722.

(34) Gertz, M. A.; Landau, H.; Comenzo, R. L.; Seldin, D.; Weiss, B.; Zonder, J.; Merlini, G.; Schonland, S.; Walling, J.; Kinney, G. G.; Koller, M.; Schenk, D. B.; Guthrie, S. D.; Liedtke, M. First-in-Human Phase I/II Study of NEOD001 in Patients With Light Chain Amyloidosis and Persistent Organ Dysfunction. *J. Clin. Oncol.* **2016**, *34*, 1097–1103.

(35) Nuvolone, M.; Merlini, G. Systemic amyloidosis: novel therapies and role of biomarkers. *Nephrol. Dial. Transplant.* **2017**, *32*, 770–780.

(36) Velander, P.; Wu, L.; Henderson, F.; Zhang, S.; Bevan, D. R.; Xu, B. Natural product-based amyloid inhibitors. *Biochem. Pharmacol.* **2017**, *139*, 40–55.

(37) Baden, E. M.; Owen, B. A.; Peterson, F. C.; Volkman, B. F.; Ramirez-Alvarado, M.; Thompson, J. R. Altered dimer interface decreases stability in an amyloidogenic protein. *J. Biol. Chem.* **2008**, *283*, 15853–15860.

(38) Blancas-Mejía, L. M.; Tischer, A.; Thompson, J. R.; Tai, J.; Wang, L.; Auton, M.; Ramirez-Alvarado, M. Kinetic control in protein folding for light chain amyloidosis and the differential effects of somatic mutations. *J. Mol. Biol.* **2014**, *426*, 347–361.

(39) Piehl, D. W.; Blancas-Mejia, L. M.; Ramirez-Alvarado, M.; Rienstra, C. M. Solid-state NMR chemical shift assignments for AL-09 VL immunoglobulin light chain fibrils. *Biomol. NMR Assign.* **2017**, *11*, 45–50.

(40) Piehl, D. W.; Blancas-Mejia, L. M.; Wall, J. S.; Kennel, S. J.; Ramirez-Alvarado, M.; Rienstra, C. M. Immunoglobulin light chains form an extensive and highly ordered fibril involving the N- and C-termini. *ACS Omega* **2017**, *2*, 712–720.

(41) Hohwy, M.; Jaroniec, C. P.; Reif, B.; Rienstra, C. M.; Griffin, R. G. Local structure and relaxation in solid-state NMR: accurate measurement of amide N-H bond lengths and H-N-H bond angles. *J. Am. Chem. Soc.* **2000**, *122*, 3218–3219.

(42) Rienstra, C. M.; Hohwy, M.; Mueller, L. J.; Jaroniec, C. P.; Reif, B.; Griffin, R. G. Determination of multiple torsion-angle constraints in U–13C,15N-labeled peptides: 3D 1H–15N–13C–1H dipolar chemical shift NMR spectroscopy in rotating solids. *J. Am. Chem. Soc.* **2002**, *124*, 11908–11922.

(43) Zhao, X.; Eden, M.; Levitt, M. H. Recoupling of heteronuclear dipolar interactions in solid-state NMR using symmetry-based pulse sequences. *Chem. Phys. Lett.* **2001**, *342*, 353–361.

(44) Franks, W. T.; Kloepper, K. D.; Wylie, B. J.; Rienstra, C. M. Four-dimensional heteronuclear correlation experiments for chemical shift assignment of solid proteins. *J. Biomol. NMR* **2007**, *39*, 107–131.

(45) Li, Y.; Wylie, B. J.; Rienstra, C. M. Selective refocusing pulses in magic-angle spinning NMR: characterization and applications to multi-dimensional protein spectroscopy. *J. Magn. Reson.* **2006**, *179*, 206–216.

(46) Ying, J.; Delaglio, F.; Torchia, D. A.; Bax, A. Sparse multidimensional iterative lineshape-enhanced (SMILE) reconstruction of both non-uniformly sampled and conventional NMR data. *J. Biomol. NMR* **2017**, *68*, 101–118.

(47) Hoch, J. C.; Maciejewski, M. W.; Mobli, M.; Schuyler, A. D.; Stern, A. S. Nonuniform sampling and maximum entropy reconstruction in multidimensional NMR. *Acc. Chem. Res.* **2014**, *47*, 708–717.

(48) Palmer, M. R.; Suiter, C. L.; Henry, G. E.; Rovnyak, J.; Hoch, J. C.; Polenova, T.; Rovnyak, D. Sensitivity of nonuniform sampling NMR. *J. Phys. Chem. B* **2015**, *119*, 6502–6515.

(49) Tycko, R.; Hu, K. N. A Monte Carlo/simulated annealing algorithm for sequential resonance assignment in solid state NMR of uniformly labeled proteins with magic-angle spinning. *J. Magn. Reson.* **2010**, *205*, 304–314.

(50) Hu, K. N.; Qiang, W.; Tycko, R. A general Monte Carlo/simulated annealing algorithm for resonance assignment in NMR of uniformly labeled biopolymers. *J. Biomol. NMR* **2011**, *50*, 267–276.

(51) Yang, Y.; Fritzsching, K. J.; Hong, M. Resonance assignment of the NMR spectra of disordered proteins using a multi-objective non-dominated sorting genetic algorithm. *J. Biomol. NMR* **2013**, *57*, 281–296.

CHAPTER 2: Solid-state NMR Chemical Shift Assignments for AL-09 V_L Immunoglobulin Light Chain Fibrils

2.1. NOTES AND ACKNOWLEDGEMENTS

This chapter is adapted with permission from Springer Science and Business Media: *Biomolecular NMR Assignments*. Solid-state NMR chemical shift assignments for AL-09 V_L immunoglobulin light chain fibrils, volume 11, 2017, pages 45-50, authors: Dennis W. Piehl*, Luis M. Blancas-Mejía*, Marina Ramirez-Alvarado, Chad M. Rienstra. *These authors contributed equally to this work. This research is supported by the University of Illinois (Centennial Scholars Award to C.M.R.), R01-GM075514 (to M.R.A.), the Mayo Foundation, and the generous support of amyloidosis patients and their families. D.W.P. is an American Heart Association Predoctoral Fellow (15PRE25100008). We thank Marcus D. Tuttle and Alexander M. Barclay for help with SSNMR data acquisition and processing.

2.2. ABSTRACT

Light chain (AL) amyloidosis is a systemic disease characterized by the formation of immunoglobulin light-chain fibrils in critical organs of the body. The light-chain protein AL-09 presents one severe case of cardiac AL amyloidosis, which contains seven mutations in the variable domain (V_L) relative to its germline counterpart, κI O18/O8 V_L. Three of these mutations are non-conservative—Y87H, N34I, and K42Q—and previous work has shown that they are responsible for significantly reducing the protein's thermodynamic stability, allowing fibril formation to occur with fast kinetics and across a wide-range of pH conditions. Currently, however, there is extremely limited structural information available which explicitly describes the residues that are involved in supporting the misfolded fibril structure. Here, we assign the site-specific ^{15}N and ^{13}C chemical shifts of the rigid core residues of AL-09 V_L fibrils by solid-state NMR, reporting on the regions of the protein involved in the fibril as well as the extent of secondary structure.

2.3. BIOLOGICAL CONTEXT

Light chain (AL) amyloidosis is an acute systemic disease in humans that is described by the formation of immunoglobulin light-chain fibrils in the extracellular matrix of critical organs of the body[1]. AL patients typically die within two to three years of diagnosis, and this expectancy is reduced to under a year if the affected organ is the heart—which comprises approximately 50 % of all AL cases[2-4]. Currently, there are only a limited number of therapies that exist which can marginally extend this outlook[1,4]. The light chains that compose these fibrils originate from abnormally proliferative monoclonal plasma cells, which secrete a large amount of the amyloidogenic protein into the bloodstream[5]. Although these proteins are natively innocuous, they can become toxic in the blood upon misfolding, providing the nucleation event for fibril elongation and deposition to occur[1]. Although the exact reason for the misfolding of these proteins is not completely understood, studies on the thermodynamic stability of patient-derived light chains reveal that the introduction of particular somatic (non-inherited) mutations within the V_L of the germline sequence can predispose them to forming amyloidogenic fibrils[1,6].

One particularly severe case of AL arises from the protein AL-09 (from patient AL-09)[7]. The AL-09 V_L domain is a 108 amino acid residue protein composed of primarily β-sheet secondary structure that contains seven mutations relative to its non-amyloidogenic germline counterpart, κI O18/O8 (Figure 2.1a–b)[1,6,8]. Notably, AL-09 is observed to form fibrils spontaneously and with fast kinetics at a physiological pH of 7.4, whereas κI O18/O8 requires the addition of a pre-formed fibril to initiate the aggregation reaction (yet, both proteins spontaneously form fibrils at a pH of 2.0) (Figure 2.1c)[5-7]. Interestingly, restorative mutational analyses for AL-09 have revealed that the three non-conservative mutations are responsible for a significant amount of the destabilizing effect—Y87H, N34I, and K42Q (in decreasing order of influence)[1,5,8]. However, due to a lack of high-resolution structural or chemical shift information of the fibrillar species, the exact influence that these mutations are suspected to have on increasing the fibril-forming propensity of AL-09 relative to κI O18/O8 remains unclear[1,5]. Accordingly, in this study, the sequence composition and secondary structure of the rigid regions for AL-09 V_L fibrils

were investigated by magic-angle spinning (MAS) SSNMR spectroscopy, revealing an extensive and highly ordered fibril core that involves residues of both the N- and C-terminus.

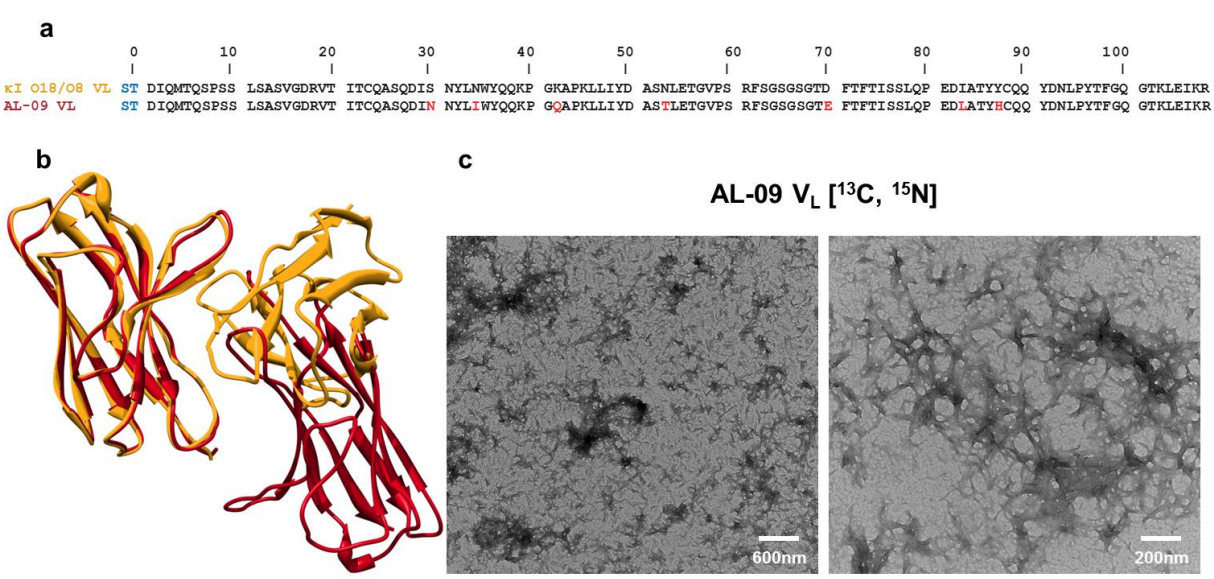

Figure 2.1. a. Sequence alignment for AL-09 and κI O18/O8 V_L (residues in red indicate mutations; those in blue are purification artifacts). **b.** X-ray structures of AL-09 and κI V_L dimers in their native forms (PDBs: 2Q1E and 2Q20, respectively)[8]. **c.** TEM images of [^{13}C, ^{15}N]-labeled AL-09 fibrils used for SSNMR studies.

2.4. METHODS AND EXPERIMENTS

Protein Expression and Purification. AL-09 V_L is a patient-derived variable domain proteins belonging to the κI O18/O8 germline gene product (also known as IGKV 1-33)[2]. Uniformly [^{13}C, ^{15}N]-labeled AL-09 V_L protein was expressed using a modified version of the protocol described by Marley et al. 2001. Briefly, *Escherichia coli* BL21 (DE3) Gold competent cells (Stratagene, La Jolla, CA) were transformed with pET12a-AL-09 V_L and grown in 1 L of 2XYT media at 37 ºC and 250 rpm. Upon reaching an A_{600nm} of 0.6, cells were pelleted for 20 min at 4,000 *g*, and subsequently washed and re-pelleted (20 min, 4,000 *g*) using an M9 salt solution[10] that excluded all nitrogen and carbon sources. The cell pellet was resuspended in 250 mL of isotopically labeled M9 minimal medium containing [^{13}C]-

glucose and [^{15}N]-ammonium chloride (Cambridge Isotope Laboratories, Tewksbury, MA) as the sole carbon and nitrogen sources, and incubated for 1 h at 37 ºC and 250 rpm to allow for recovery of growth. Cells were induced by adding isopropyl β-D-1-thiogalactopyranoside (IPTG) for a concentration of 0.8 mM and then grown at 25 °C and 100 rpm for 12 h before harvesting by centrifugation at 8,000 rpm for 20 min.

Extraction of uniformly [^{13}C, ^{15}N]-labeled AL-09 V_L protein was accomplished by osmotic shock, as follows. Harvested cell pellets were resuspended in 100 mL (each) of cold phosphate-buffered saline (PBS), incubated on ice for 10 min, and pelleted for 20 min at 4,000 *g*. Subsequently, cell pellets were resuspended in 100 mL of 20% sucrose solution, incubated on ice for 10 min, and pelleted for 20 min at 8,000 *g*. Upon resuspension in 100 mL of 10 mM Tris–HCl buffer, pH 7.4, each pellet was sonicated three times for 10 s at 35% amplitude (with 1 min intervals of incubation on ice), using a Misonix S-400 sonicator (Osonica, Newtown, CT). Sonicated cells were pelleted for 20 min at 8,000 *g*, followed by resuspension in 100 mL of 5 M urea, incubation on ice for 30 min, and re-centrifugation for 1 h at 20,000 *g*. The resulting pellets were discarded and the urea supernatant was dialyzed overnight against 4 L (per pellet) of 10 mM Tris–HCl buffer, pH 7.4. Uniformly [^{13}C, ^{15}N]-labeled AL-09 V_L monomer was purified under the same buffer conditions by size exclusion chromatography with a HiLoad 16/60 Superdex 75 column on an AKTA FPLC system (GE Healthcare, Piscataway, NJ).

Protein purity and homogeneity were determined by SDS-PAGE. Protein concentration was determined by UV absorption at 280 nm using an extinction coefficient of ε = 13,610 M^{-1} cm^{-1} for AL-09 V_L, calculated from the amino acid sequence. Pure protein was flash frozen and stored at −80 °C. As a quality control measurement, far UV circular dichroism (CD) spectroscopy was used to confirm that the protein retained the native secondary structure previously reported at the beginning of the fibril formation reaction. Far UV-CD spectra from 260–200 nm (1 nm bandwidth) were acquired at 4 °C, on a Jasco Spectropolarimeter 810 (JASCO Inc., Easton, MD) using a 0.2 cm path-length quartz cuvette. All samples were prepared with 20 μM protein in 10 mM Tris–HCl pH 7.4.

Mass Spectrometry. The molecular mass of uniformly [^{13}C, ^{15}N]-labeled AL-09 V_L protein was determined using MALDI-TOF. The average molecular weight was measured to be 12776 Da, indicating ~99 % [^{13}C, ^{15}N] incorporation relative to the calculated masses for unlabeled and 100 % [^{13}C, ^{15}N]-labeled protein (11933 Da and 12789 Da, respectively).

***In vitro* Fibril Formation.** Uniformly [^{13}C, ^{15}N]-labeled AL-09 V_L fibril samples were prepared from protein monomer solutions, according to previously reported protocols[6]. Briefly, purified protein was thawed at 4 ºC, filtered using 0.45 μm membranes and subjected to ultracentrifugation at a speed of 90,000 rpm (645,000 *g*) for 3.3 h in a NVT-90 rotor on an Optima L-100 XP centrifuge (Beckman Coulter, Brea, CA), to remove any pre-formed aggregates formed during the thawing of the soluble protein.

Ultracentrifuged protein samples were prepared on ice at a final concentration of 20 μM in 10 mM Acetate Borate Citrate (ABC) buffer pH 2.0, containing 150 mM NaCl, 10 μM Thioflavin T (ThT), and 0.02 % NaN_3 in 1.5 mL low binding microcentrifuge tubes (1.0 mL total volume). Fibril formation reactions were conducted in triplicate using sealed, black 96-well polystyrene plates (Nunc, Roskilde, Denmark; Greiner, Monroe, NC) with 260 μL reaction mixture per well and incubating at 37 °C with continuous orbital shaking (300 rpm) in a New Brunswick Scientific Innova40 incubator shaker until fibril formation was complete. Fibrillation progress was monitored by the change in enhanced fluorescence intensity of ThT, and the presence of fibrils was subsequently confirmed by transmission electron microscopy (TEM). On average, the time required for fibrillation to conclude for AL-09 V_L fibrils is ~2 days (50 hours).

Transmission Electron Microscopy (TEM). To confirm the presence of fibrils, a 3 μL fibril sample was placed on a 300 mesh copper formvar/carbon grid (Electron Microscopy Science, Hatfield, PA), and excess liquid was removed. The samples were negatively stained with 2 % uranyl acetate, washed twice with H_2O, and air-dried. Grids were analyzed on a Philips Tecnai T12 transmission electron microscope at 80 kV (FEI, Hillsboro, OR).

Preparation of Fibrils for SSNMR Studies. Fibril samples were transferred to polypropylene microfuge tubes (Item #: 357448) and pelleted in a table-top ultracentrifuge with a TLA-100.3 (Beckman Coulter) rotor and microfuge tube adapters (Item #: 355919), for 1 hr at 4 ºC and 55,000 rpm (or ~100,000 *g*, accounting for the shorter radius of the microfuge tubes). Following the removal of the fibrillation buffer, fibril pellets were washed twice with deionized water using homogenization and subsequent ultracentrifugation. Washed fibril pellets were dried completely under nitrogen gas by monitoring the mass of the sample until it remained unchanged. Dried AL-09 V_L fibril samples were packed into either 3.2 mm (outer-diameter) standard-wall rotor or 3.2 mm thin-wall rotor (Agilent Technologies, Santa Clara, CA), and hydrated to a level of 50 % water by mass. Kel-F and rubber spacers were used to retain hydration level of sample.

Solid-State NMR Spectroscopy. MAS SSNMR spectra of AL-09 V_L fibrils were acquired at 11.7 T (500 MHz ^{1}H frequency) and 17.6 T (750 MHz) on VNMRS (Agilent Technologies) spectrometers with a 3.2 mm Balun or BioMAS ^{1}H-^{13}C-^{15}N probe as well as at 14.1 T (600 MHz) on an InfinityPlus spectrometer using a 3.2 mm Balun or T3 probe. For all multidimensional experiments, the MAS rate was set to 11.111 kHz (500 MHz), 12.5 kHz (750 MHz), or 13.333 kHz (600 MHz), and a variable-temperature (VT) stack was used to provide a constant flow of -10 ºC, 0 ºC, or 10 ºC air, for a sample temperature of -5 to 15 ºC (due to heating caused by MAS and 65–80 kHz ^{1}H SPINAL-64[11] decoupling power). Multidimensional SSNMR experiments performed on AL-09 V_L fibril samples include: ^{13}C-^{13}C (CC) 2D, ^{15}N-^{13}Cα (NCA) 2D, ^{15}N-^{13}C' (NCO) 2D, ^{15}N-(^{13}Cα)-^{13}CX (N(CA)CX) 2D, ^{15}N-(^{13}C')-^{13}CX (N(CO)CX) 2D, ^{15}N-^{13}Cα-^{13}CX (NCACX) 3D, ^{15}N-^{13}C'-^{13}CX (NCOCX) 3D, ^{13}Cα-^{15}N-^{13}C' (CANCO) 3D, ^{13}Cα-^{15}N-(^{13}C')-^{13}CX (CAN(CO)CX) 3D, and ^{13}C-^{13}C-^{13}C (CCC) 3D. One-bond ^{13}C-^{13}C correlations were achieved using either dipolar-assisted rotational resonance (DARR; τ_{mix} = 25–90 ms)[12] or supercycled POST-C7 (SPC-7; τ_{mix} = 0.9 ms or 1.2 ms)[13] for magnetization transfer, and one-bond ^{15}N-^{13}C or ^{13}C-^{15}N correlations were accomplished using SPECIFIC-CP[14]. Further experimental details are summarized in Table 2.1.

Solid-State NMR Data Processing and Analysis. All SSNMR data were processed with forward or backward linear prediction, zero filling, and Lorentzian-to-Gaussian apodization using NMRPipe[15]. Peak assignments, signal-to-noise ratio (SNR) measurements, and linewidth measurements were performed using Sparky[16]. Secondary structure analysis of site-specifically resolved residues was performed using TALOS-N[17].

2.5. ASSIGNMENTS AND DATA DEPOSITION

The number of residues involved in the structured region of uniformly [^{13}C, ^{15}N]-labeled AL-09 V_L fibrils was assessed by the acquisition and analysis of multidimensional MAS SSNMR spectra (Figure 2.2). Specifically, a count of the number of peaks observed in initial 2D ^{13}C-^{13}C (CC) spectra allowed for an approximation of >50 spin-systems at a signal threshold of 10 times the noise floor, σ. This process was facilitated by the focus on Ala and Thr spin-systems, which exhibit highly resolved side-chain peaks in ^{13}C-^{13}C 2D spectra (Figure 2.2b–c). This estimate was further improved by repeating the process for higher resolution 3D data sets—specifically, NCACX, NCOCX, and CAN(CO)CX spectra (at a signal threshold of 8σ)—which consistently revealed the presence of 70 ± 15 spin-systems. Notably, as there are five alanine residues in the AL-09 V_L sequence, the observation of no more than five alanine ^{13}Cα-^{13}Cβ peaks in the CC 2D (Figure 2.2b) and 3D spectra provided initial support for the existence of a homogenous fibril conformation. Furthermore, the majority of the resolved cross-peaks between the α- and β-carbons in each residue are indicative of a primarily β-sheet secondary structure (Figure 2.2a–c). Together, these data reveal that approximately 70 residues of the 110 total in the AL-09 sequence are observed in the rigid fibril structure.

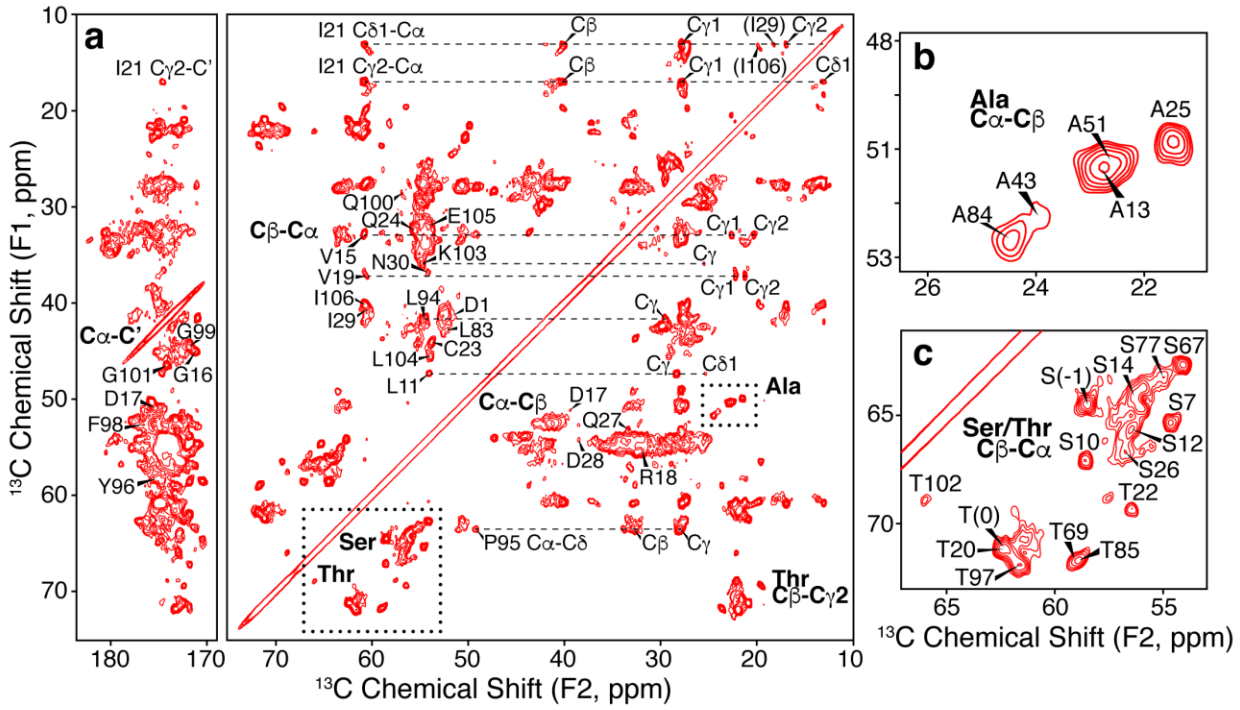

Figure 2.2. a. ^{13}C-^{13}C 2D SSNMR spectra of [^{13}C, ^{15}N]-labeled AL-09 V_L fibrils collected at 750 MHz and 0 ºC using 40 ms ^{13}C-^{13}C DARR (dipolar) mixing and 12.5 kHz MAS. Labels indicate a subset of site-specific residue assignments and spectral regions unique to specific amino acid types. **b.** Expansions of the Ala ^{13}Cα-^{13}Cβ region and **c.** Ser/Thr ^{13}Cβ-^{13}Cα region illustrate the extent of assignments and predominantly β-sheet secondary chemical shift trends.

To pursue the site-specific identification of the residues involved in the highly structured fibril region, chemical shift assignments were made via a backbone-walk using a set of NCACX, NCOCX, and CAN(CO)CX 3D spectra (experimental details for each are described in Table 2.1). This analysis enabled the complete backbone assignments of 48 residues site-specifically within the AL-09 sequence, as well as partial assignments for an additional 8 residues (Figure 2.3 and 2.4). Furthermore, we have also made tentative assignments for residues I34, R61–G66, and S76–S77 (9 total); however, because the sensitivity of these spin-systems was relatively limited in the 3D data sets we could not confidently confirm the connectivity between these residues and their neighbors, and thus opted to exclude these assignments here. Interestingly, the majority of the assigned resonances are located within the first 35 and last 20

amino acids of the N- and C-termini, respectively. In addition, TALOS-N predictions of the secondary structure and order parameters (based on the chemical shift values of the assignments) support the existence of a highly ordered and β-sheet rich fibril structure (Figure 2.3)[17].

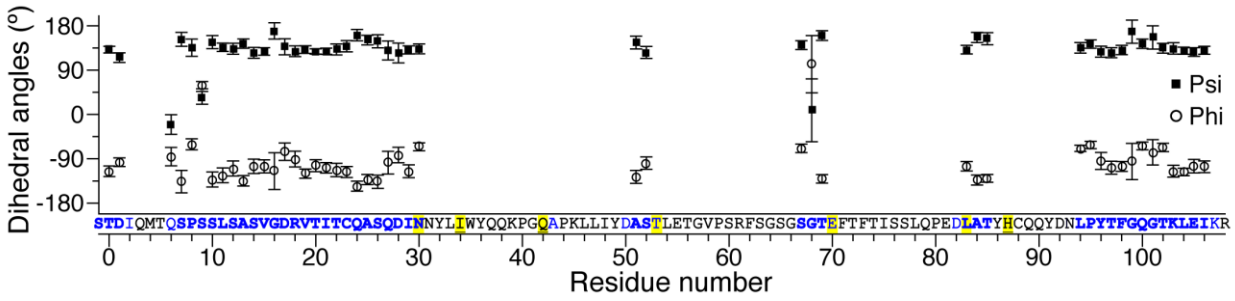

Figure 2.3. Secondary structure and extent of chemical shift assignments for AL-09 V_L fibrils. TALOS-N predicted backbone dihedral angles psi (closed squares) and phi (open circles) are plotted a function of residue number; error bars correspond to the standard deviation of the averaged TALOS-N matches. AL-09 residues for which chemical shifts have been assigned are labeled in blue; those in bold are unambiguous and have resonances assigned to all three backbone atoms. Highlighted residues indicate the location at which mutations occur; those that are underlined are non-conservative

Overall, our results reveal that approximately 70 of the 110 residues in the AL-09 V_L sequence are involved in the highly ordered fibril with β-strand secondary structure. Moreover, we observed that the majority of these rigid residues are located in the N- and C-termini of the AL-09 V_L sequence and generally demonstrate strong and well-resolved backbone signals. By contrast, the residues located between the two termini (i.e., those immediately following N30 and preceding L94) exhibit relatively weaker signal intensities and/or chemical shift degeneracy, consequently leading to ambiguity. Nevertheless, the majority of these spin-systems were able to be assigned by amino acid type based on comparison with empirical chemical shift statistics. Further, the cross-peak intensities exhibited by these spin-systems were observed in most cases to range from 60–130% relative to the average peak intensity of site-specifically assigned residues. Together these observations provide support for the existence of a

single conformation of the AL-09 V_L fibrils, as the distribution of total intensity between several different conformations of a single spin-system would lead to a substantial reduction in the intensity of each cross-peak. Finally, one interesting correlation underlying the extent of chemical shift assignments reported here is that all seven mutations are positioned in the central portion of the AL-09 V_L sequence, and only two of which have assignments for all backbone atoms (N30 and L83). Although a detailed understanding of the interactions engaged by the non-conservative mutations within the fibril structure will require completion of all chemical shift assignments, it is nonetheless noteworthy that the mutations are not the sole constituents of the ordered fibril core.

2.6. ACCESSION NUMBERS

Chemical shift assignments for AL-09 V_L fibrils can be accessed on the BioMagResBank (BMRB) under entry number 26879. The GenBank accession number for the light chain cDNA of AL-09 V_L is AF490909.

2.7. SUPPORTING INFORMATION

Figure 2.4. Strip-plot of all unambiguously assigned residues for [^{13}C, ^{15}N]-labeled AL-09 V_L fibrils. Sequential backbone and/or side-chain chemical shift assignments are illustrated using 3D NCACX (red), NCA(CB)CO (positive peaks, red; negative peaks, green), NCOCX (blue), and CANcoCX (black) spectra for residues **a.** S(-1)–D1, and S7–N30, as well as **b.** A51–S52, S67–T69, L83–T85, and L94–I106. Strips from the NCA(CB)CO 3D data were used in place of NCACX 3D strips for instances when the latter experiment did not provide unambiguous cross-peak data for the corresponding assignment.

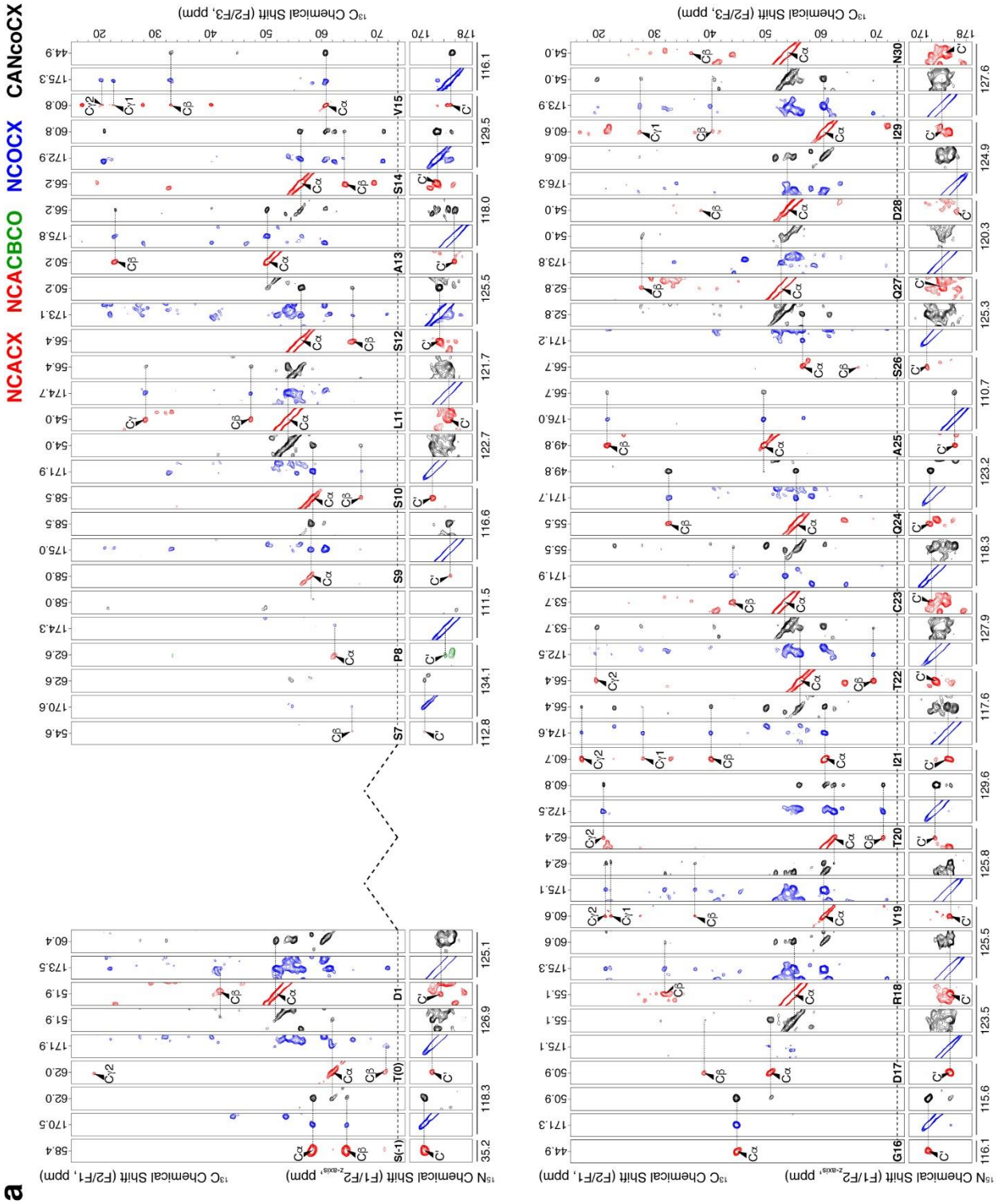

Figure 2.4. (cont.) a. S(-1)–D1, and S7–N30

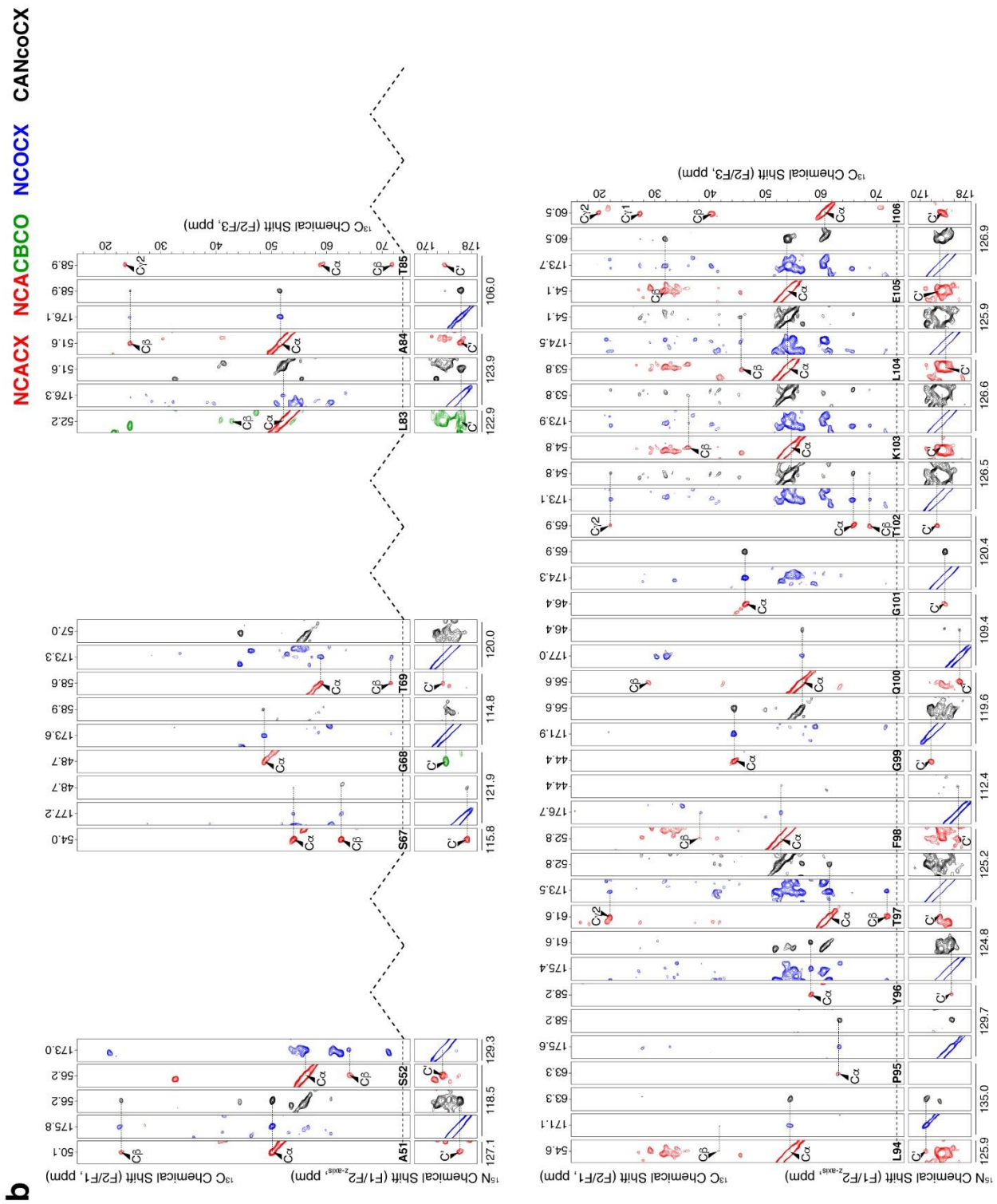

Figure 2.4. (cont.) b. A51–S52, S67–T69, L83–T85, and L94–I106

Table 2.1. Experimental details for SSNMR data collected on [^{13}C,^{15}N]-labeled AL-09 V_L fibrils used to enable chemical shift assignments.

Experiment	**CC**	**NCA**	**NCO**	**NCACX**		**NCA(CB)CO**	**NCOCX**		**NCOCA**	**CANCO**	**CAN(CO)CX**	**CCC**
Spectrometer ^{1}H frequency	750 MHz	600 MHz	600 MHz	600 MHz	500 MHz	600 MHz	600 MHz	500 MHz	600 MHz	600 MHz	600 MHz	600 MHz
Probe	3.2 mm Balun	3.2 mm Balun	3.2 mm Balun	3.2 mm Balun	3.2 mm Balun	3.2 mm Balun	3.2 mm Balun	3.2 mm Balun	3.2 mm Balun	3.2 mm Balun	3.2 mm Balun	3.2 mm T3
MAS (kHz)	12.500	13.333	13.333	13.333	11.111	13.333	13.333	11.111	13.333	13.333	13.333	13.333
Temperature set-point (ºC)	0.0	-10.0	-10.0	-10.0	10.0	-10.0	-10.0	10.0	-10.0	-10.0	-10.0	-10.0
Measurement time (h)	42.6	3.48	3.48	112	120	126	112	90	70	84	168	168
Number of scans/row	8	8	8	4	4	8	4	4	8	4	8	4
Pulse delay (s)	2.0	2.0	2.0	2.0	1.5	2.0	2.0	1.5	2.0	2.0	2.0	1.0
Polarization transfer 1	H-C CP	H-N CP	H-N CP	H-N CP	H-N CP	H-N CP	H-N CP	H-N CP	H-N CP	H-N CP	H-N CP	H-C CP
Transfer time (ms)	1.2	1.5	1.5	1.5	2.0	1.5	1.5	2.0	1.5	1.0	1.0	1.0
RF power (kHz)	47(H)/77(C)	69(H)/55(N)	67(H)/55(N)	74(H)/52(N)	61(H)/60(N)	68(H)/55(N)	71(H)/52(N)	61(H)/60(N)	67(H)/55(N)	78(H)/53(CA)	78(H)/53(CA)	83(H)/74(C)
Shape	tangent	tangent	tangent	tangent	tangent	tangent	tangent	tangent	tangent	tangent	tangent	tangent
Carrier (ppm)	97.5	117.5	117.5	117.5	83.3	117.5	117.5	100.1	117.5	57.5	57.5	80.0
Polarization transfer 2	DARR	N-CA CP	N-CO CP	N-CA CP	N-CA CP	N-CA CP	N-CO CP	N-CO CP	N-CO CP	CA-N CP	CA-N CP	SPC-7
Transfer time (ms)	40.0	5.0	5.0	4.0	4.5	5.0	4.0	3.0	5.0	4.0	4.0	0.9
RF power (kHz)	10.6 (H)	32(N)/22(CA)	35(N)/51(CO)	31(N)/21(CA)	28(N)/15(CA)	32(N)/21(CA)	31(N)/48(CO)	28(N)/14(CO)	35(N)/52(CO)	21(CA)/31(N)	21(CA)/31(N)	95 (C)
Shape	-	tangent	tangent	tangent	tangent	tangent	tangent	tangent	tangent	tangent	tangent	-
Carrier (ppm)	-	58.0	175.0	57.5	44.0	58.0	174.0	152.4	175.0	115.0	115.0	-
Polarization transfer 3				DARR	DARR	SPC-7	DARR	DARR	SPC-7	N-CO CP	N-CO CP	DARR
Transfer time (ms)				25.0	90.0	1.2	25.0	90.0	1.2	6.0	6.0	50.0
RF power (kHz)				17 (H)	13 (H)	98 (C)	17 (H)	13 (H)	92 (C)	31(N)/44(CO)	31(N)/45(CO)	18 (H)
Shape				-	-	-	-	-	-	tangent	tangent	-
Carrier (ppm)				-	-	-	-	-	-	57.5	57.5	-
Polarization transfer 4											DARR	
Transfer time (ms)											25.0	
RF power (kHz)											18 (H)	
Shape											-	
Carrier (ppm)											-	
Increments in ω1	512	256	256	128	256	64	128	96	48	128	128	192
Sweep width (kHz)	5.000	6.666	6.666	3.333	11.111	2.222	3.333	5.555	3.333	6.666	6.666	13.333
Evol. time (ms)	10.24	19.20	19.20	19.20	23.04	14.40	19.20	17.28	7.20	9.60	9.60	7.20
Increments in ω2	2000	1024	1024	96	64	64	96	64	48	96	96	192
Sweep width (kHz)	100.000	66.667	66.667	4.444	11.111	6.666	4.444	11.111	3.333	3.333	3.333	13.333
Evol. or Acq. time (ms)	20.00	15.36	15.36	10.80	5.76	4.80	10.80	5.76	7.20	14.40	14.40	7.20
Increments in ω3				1024	2000	1024	1024	2000	1024	1024	1024	1360
Sweep width (kHz)				66.667	100.000	66.667	66.667	100.000	66.667	66.667	66.667	80.000
Acq. time (ms)				15.36	20.00	15.36	15.36	20.00	15.36	15.36	15.36	17.00
Decoupling	SPINAL64	SPINAL64	SPINAL64	SPINAL64	SPINAL64	SPINAL64	SPINAL64	SPINAL64	SPINAL64	SPINAL64	SPINAL64	SPINAL64
RF power (kHz)	70.4	77.8	77.8	77.4	83.9	77.8	77.4	83.9	77.8	77.4	77.4	77.4

2.8. REFERENCES

(1) Baden, E. M.; Sikkink, L. A.; Ramirez-Alvarado, M. Light chain amyloidosis – current findings and future prospects. *Curr. Protein Pept. Sci.* **2009**, *10*, 500–508.

(2) Abraham, R. S.; Geyer, S. M.; Price-Troska, T. L.; Allmer, C.; Kyle, R. A.; Gertz, M. A.; Fonseca, R. Immunoglobulin light chain variable (V) region genes influence clinical presentation and outcome in light chain-associated amyloidosis (AL). *Blood* **2003**, *101*, 3801–3808.

(3) Levinson, R. T.; Olatoye, O. O.; Randles, E. G.; Howell, K. G.; DiCostanzo, A. C.; Ramirez-Alvarado, M. Role of mutations in the cellular internalization of amyloidogenic light chains into cardiomyocytes. *Sci. Rep.* **2013**, *3*, 1–8.

(4) McLaughlin, R. W.; De Stigter, J. K.; Sikkink, L. A.; Baden, E. M.; Ramirez-Alvarado, M. The effects of sodium sulfate, glycosaminoglycans, and Congo red on the structure, stability, and amyloid formation of an immunoglobulin light-chain protein. *Protein Sci.* **2006**, *15*, 1710–1722.

(5) Ramirez-Alvarado, M. Amyloid formation in light chain amyloidosis. *Curr. Top. Med. Chem.* **2012**, *12*, 2523–2533.

(6) Blancas-Mejía, L. M.; Tischer, A.; Thompson, J. R.; Tai, J.; Wang, L.; Auton, M.; Ramirez-Alvarado, M. Kinetic control in protein folding for light chain amyloidosis and the differential effects of somatic mutations. *J. Mol. Biol.* **2014**, *426*, 347–361.

(7) Martin, D. J.; Ramirez-Alvarado, M. Comparison of amyloid fibril formation by two closely related immunoglobulin light chain variable domains. *Amyloid* **2010**, *17*, 129–136.

(8) Baden, E. M.; Owen, B. A.; Peterson, F. C.; Volkman, B. F.; Ramirez-Alvarado, M.; Thompson, J. R. Altered dimer interface decreases stability in an amyloidogenic protein. *J. Biol. Chem.* **2008**, *283*, 15853–15860.

(9) Marley, J.; Lu, M.; Bracken, C. A method for efficient isotopic labeling of recombinant proteins. *J. Biomol. NMR* **2001**, *20*, 71–75.

(10) Sambrook, J.; Russell, D. W.; Irwin, N.; Janssen, K. A. In *Molecular Cloning: A Laboratory Manual*; 3 ed.; Argentine, J., Ed.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York, 2001; Vol. 3, pp A2.2.

(11) Comellas, G.; Lopez, J. J.; Nieuwkoop, A. J.; Lemkau, L. R.; Rienstra, C. M. Straightforward, effective calibration of SPINAL-64 decoupling results in the enhancement of sensitivity and resolution of biomolecular solid-state NMR. *J. Magn. Reson.* **2011**, *209*, 131–135.

(12) Takegoshi, K.; Nakamura, S.; Terao, T. 13C-1H dipolar-assisted rotational resonance in magic-angle spinning NMR. *Chem. Phys. Lett.* **2001**, *344*, 631–637.

(13) Hohwy, M.; Rienstra, C. M.; Jaroniec, C. P.; Griffin, R. G. Fivefold symmetric homonuclear dipolar recoupling in rotating solids: Application to double quantum spectroscopy. *J. Chem. Phys.* **1999**, *110*, 7983–7992.

(14) Baldus, M.; Petkova, A. T.; Herzfeld, J.; Griffin, R. G. Cross polarization in the tilted frame: assignment and spectral simplification in heteronuclear spin systems. *Mol. Phys.* **1998**, *95*, 1197–1207.

(15) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. NMRPipe: a multidimensional spectral processing system based on unix pipes. *J. Biomol. NMR* **1995**, *6*, 277–293.

(16) Goddard, T. D.; Kneller, D. G. SPARKY 3. University of California, San Francisco, **2008**.

(17) Shen, Y.; Bax, A. Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J. Biomol. NMR* **2013**, *56*, 227–241.

CHAPTER 3: Immunoglobulin Light Chains Form an Extensive and Highly Ordered Fibril Involving the N- and C-termini

3.1. NOTES AND ACKNOWLEDGEMENTS

This chapter is adapted with permission from: *ACS Omega*. Immunoglobulin light chains form an extensive and highly ordered fibril involving the N- and C-termini, volume 2, 2017, pages 712-720, authors: Dennis W. Piehl[*], Luis M. Blancas-Mejía[*], Jonathan S. Wall, Stephen J. Kennel, Marina Ramirez-Alvarado, Chad M. Rienstra. [*]These authors contributed equally to this work. This research is supported by the University of Illinois (Centennial Scholars Award to C.M.R.), R01-GM071514 (to M.R.A.), the Mayo Foundation, and the generous support of amyloidosis patients and their families. D.W.P. is an American Heart Association Predoctoral Fellow (15PRE25100008). We thank Dr. Xiangyan Shi and Joseph M. Courtney for assisting with pulse code modifications of the $R48_3^{18}$-symmetry experiment and analysis of the dipolar trajectory data, respectively.

3.2. ABSTRACT

Light chain-associated (AL) amyloidosis is a systemic disorder involving the formation and deposition of immunoglobulin light-chain fibrils in various bodily organs. One severe instance of AL disease is exhibited by the patient-derived variable domain (V_L) of the light-chain AL-09, a 108 amino acid residue protein containing seven mutations relative to the corresponding germline protein, κI O18/O8 V_L, from which it is derived. Previous work has demonstrated that the thermodynamic stability of native AL-09 V_L is greatly lowered by two of these mutations—Y87H and N34I—while a third mutation, K42Q, further increases in the kinetics of fibril formation. However, detailed knowledge regarding the residues that are responsible for stabilizing the misfolded fibril structure is lacking. In this study, we show using solid-state NMR spectroscopy that the majority of the AL-09 V_L sequence is immobilized in the fibrils and that the N- and C-terminal portions of the sequence are particularly well structured. Thus, AL-09 V_L forms an extensively ordered and β-strand-rich fibril structure. Furthermore, we demonstrate that the

predominantly β-sheet secondary structure and rigidity observed for *in vitro* prepared AL-09 V_L fibrils are qualitatively similar to those observed for AL fibrils extracted from post-mortem human spleen tissue, suggesting that this conformation may be representative of a common feature to AL fibrils.

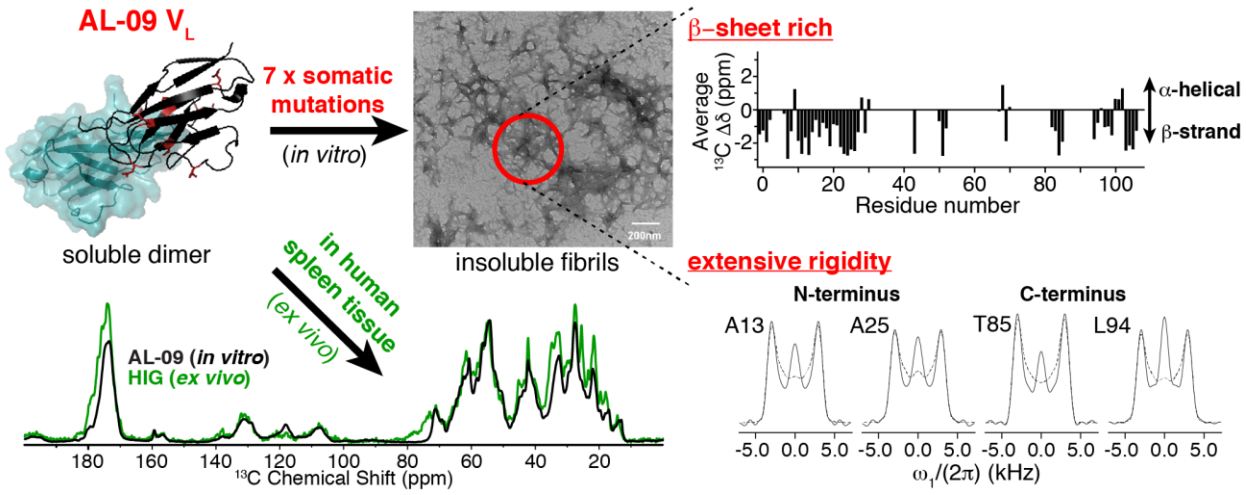

Figure 3.1. Table of contents and abstract graphic.

3.3. INTRODUCTION

Light chain (AL) amyloidosis is a severe form of systemic amyloidosis arising from the misfolding and deposition of immunoglobulin light chains as fibrils in the extracellular matrix of major organs throughout the human body, targeting sites such as the heart, kidneys, liver, spleen, and peripheral nerves.[1] The average prognosis for AL disease patients varies depending on which organ is affected, generally ranging between two to three years but dropping to below a year if the site of fibril deposition occurs at the heart.[2-4] Further, recent work has shown that the accumulation of AL fibrils around cardiomyocytes cultured *in vitro* may lead to cellular internalization of the fibrils and subsequent cell death by interrupting cell growth,[5] possibly through affecting normal metabolic function.[6] Devastatingly, the number of cardiac-inflicted AL patients accounts for approximately 50% of all AL disease cases, and despite ongoing efforts to more effectively hinder light-chain aggregation,[7-9] the availability of therapy

options remains severely limited and primarily targets the plasma cell population secreting the pathogenic protein.[1-4]

Light chain misfolding in AL disease presents a particularly interesting case of amyloidoses in that the fibril-forming light chain generally differs by a few mutations between each patient, due to the natural introduction of somatic mutations during the process of generating a wide variety of antibodies.[10] The variable domain (V_L) of light-chain protein AL-09 (from patient AL-09)[2] represents one severe instance of cardiac AL disease, which differs from its germline counterpart (κI O18/O8 V_L) by seven somatic mutations (Figure 3.2.A).[11] Three of these mutations are non-conservative—Y87H, N34I and K42Q—and are found to promote the propensity and rate of fibril formation at least in part by contributing a significant destabilizing effect on the native, soluble dimer form of AL-09 V_L.[10] In addition, previous studies on the soluble dimer structures of AL-09 V_L and κI O18/O8 V_L have demonstrated that the interface of AL-09 V_L is twisted by 90º relative to the germline dimer interface and is the site at which the three non-conservative mutations are located (Figure 3.2.C–D).[1,11] Indeed, these observations strongly suggest that the presence of these non-conservative mutations in AL-09 V_L contribute to the protein's amyloidogenicity by distorting the V_L-V_L interface of the soluble dimer structure. Further, solution-state nuclear magnetic resonance (NMR) studies on closely related V_L domains have demonstrated that the residues located at the dimer interface are predisposed to exhibiting greater conformational flexibility relative to most regions outside the interface, particularly under destabilizing conditions, thus implying that increased dynamics also contributes to the fibril forming propensity.[12,13] However, there remains limited knowledge on the extent and manner of influence that the different regions of the light-chain V_L have on stabilizing the final fibril form.

In pursuit of this interest, here we investigate the secondary structure and relative dynamics of AL-09 V_L fibrils at the residue-specific level by magic-angle spinning (MAS) solid-state NMR (SSNMR) spectroscopy. Specifically, we demonstrate that the majority of the residues in AL-09 V_L fibrils are highly rigid and exhibit a primarily β-strand secondary structure, through chemical shift and dipolar coupling

analyses. Moreover, our results reveal that most of the uniquely-assigned rigid residues are located near the N- and C-termini (i.e., before N30 and after L94) and maintain a high degree of structural order and homogeneity at a site-specific level. These observations also demonstrate notable structural differences between the fibril and soluble forms of AL-09 V_L, such as residues Y32–Q38 and K45–D50 which are ordered as a β-sheet in the dimer crystal structure (Figure 3.2.D) but are not observed site-specifically in the fibril form (Figure 3.2.A–B). By contrast, residues S12–V15 and L94–T97 are observed as β-strands in the fibril form, but fall on the outside of β-strands in the soluble form (Figure 3.2). Finally, we show that the recombinantly-generated AL-09 V_L fibrils studied here exhibit similar global structural features to those of AL fibrils directly extracted from human spleen tissue (referred to as HIG), supporting both the physiological relevance of our SSNMR studies as well as the possibility of identifying a structural motif that is common to all AL fibrils. Together, these results provide detailed insight into the extensive and unique structure of a toxic AL fibril, thus informing on the structure-based search for improved diagnostic and therapeutic strategies for AL disease.

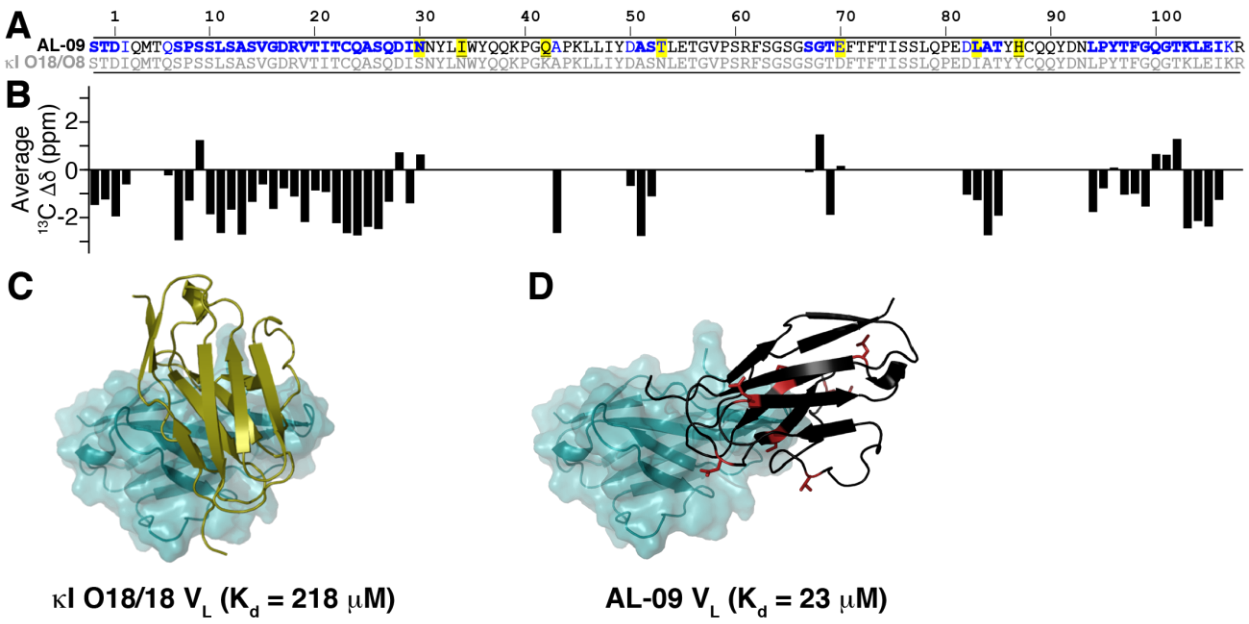

Figure 3.2. (A) Sequence alignment of AL-09 and germline κI O18/O8 V_L domains. Highlighted residues indicate mutation sites; those that are non-conservative are underlined. Residues labeled in blue are site-specifically assigned; those that are unambiguous are in bold.[14] (B) Averaged secondary ^{13}C chemical shifts, ΔδC, plotted for each assigned residue in AL-09 V_L fibrils. Values below or above the average random coil chemical shift values.[16] δC_{coil}, (i.e., $\Delta\delta C < 0$ or $\Delta\delta C > 0$ ppm) correspond to a β-strand or α-helix secondary structure, respectively. (C–D) Crystallographic structures of the soluble dimer forms of germline κI O18/O8 and AL-09 V_L domains, respectively.[11] Mutation sites in AL-09 are drawn in red.

3.4. RESULTS AND DISCUSSION

3.4.1. AL-09 V_L fibrils exhibit an extensive and β-strand rich structured region.

Based on the resonance assignments for approximately half of the protein sequence for AL-09 V_L fibrils (BMRB 26879) and observation of approximately 70 spin-systems (the majority of which have been identified at the amino acid-type specific level),[14] we pursued an analysis of the corresponding secondary structure and dynamics. In addition to obtaining initial insight on the secondary structure content of the fibril species through chemical shift-based predictions of the backbone dihedral angles using TALOS-N[15] (Figure 3.7A),[14] we performed secondary ^{13}C chemical shift (Δδ) calculations for

residues in which the $^{13}C\alpha$, $^{13}C'$, and/or $^{13}C\beta$ chemical shifts are assigned (Figure 3.2B).[16] Here we report the averaged value of $\Delta\delta$ for each residue. Collectively, these ^{13}C $\Delta\delta$ values inform on the type of secondary structure that a protein region exhibits—values close to zero indicate random coil (unstructured), while positive and negative values indicate α-helical and β-strand structure predictions, respectively.[16] As illustrated in Figure 3.2B, the large majority of uniquely assigned residues are predicted to exhibit β-sheet secondary structure, consistent with our refined TALOS-N predictions as well as findings typically observed for other fibrils.[14,17]

In addition to secondary structure information, the dipolar coupling-based SSNMR ^{13}C-^{13}C and ^{15}N-^{13}C correlation experiments report on the relative dynamics across the sequence of AL-09 V_L in its fibril state, with cross-polarization (CP) emphasizing the rigid regions. Thus, a comparison of the relative CP intensities resulting from these experiments for different residues in AL-09 V_L fibrils offers a qualitative measure of their corresponding rigidity. Here we used the relative $^{13}C\alpha$-^{15}N-$^{13}C'$ (CANCO) peak heights of assigned AL-09 V_L residues to inform on the site-to-site variation of rigidity across the sequence of the fibril species (Figure 3.7B). Our results reveal that the majority of the assigned regions of AL-09 V_L exhibit CP intensities greater than 50% of the maximum peak height for residues with non-overlapping signals. These residues are spread across the majority of the AL-09 V_L sequence. Additionally, the general increase in intensity from residue Q6 to L11 and the subsequent decrease from I21 to D28 suggests that Q6 is near the beginning of its associated β-strand while D28 is near the end, and that the flanking residues may be on the periphery of the highly ordered region.

We further assessed the relative dynamics of AL-09 V_L fibrils through a series of ^{15}N-$^{13}C\alpha$ (NCA) 2D experiments at temperatures below, near, and above the phase transition (-30 ºC, -10 ºC, and +20 ºC) of the bulk water in the samples (Figures 3.3A and 3.8). It is well known from a variety of amyloid systems that upon freezing the bulk water, CP intensity increases for those residues in mobile regions, and disordered portions of the sequence exhibit broader signals at the lower temperature.[18-20] The NCA 2D spectrum provides multiple uniquely resolved cross peaks corresponding to a wide range of residues over

the AL-09 V_L sequence, and so these experiments allow for the temperature-dependence of fibril rigidity to be assessed by monitoring the changes in signal-to-noise ratio and line width of each peak. Accordingly, the sensitivity and resolution of 11 resolved ^{15}N-$^{13}C\alpha$ cross peaks were measured at each temperature (T0, G16, D17, I21, A25, S26, S67, T85, Y96, G99, and T102). These results demonstrate a consistent number of resolvable peaks for all temperatures, with only slight increases in average line width from +20 ºC to -30 ºC (13 ± 8 Hz, ^{15}N dimension; 8 ± 19 Hz $^{13}C\alpha$ dimension) and decreases in signal-to-noise ratio (26% ± 15% lower on average). This finding indicates that the AL-09 V_L fibril structure remains highly ordered across a wide range of temperatures, and that this order is consistently maintained for the approximately 70 total spin-systems observed.[14]

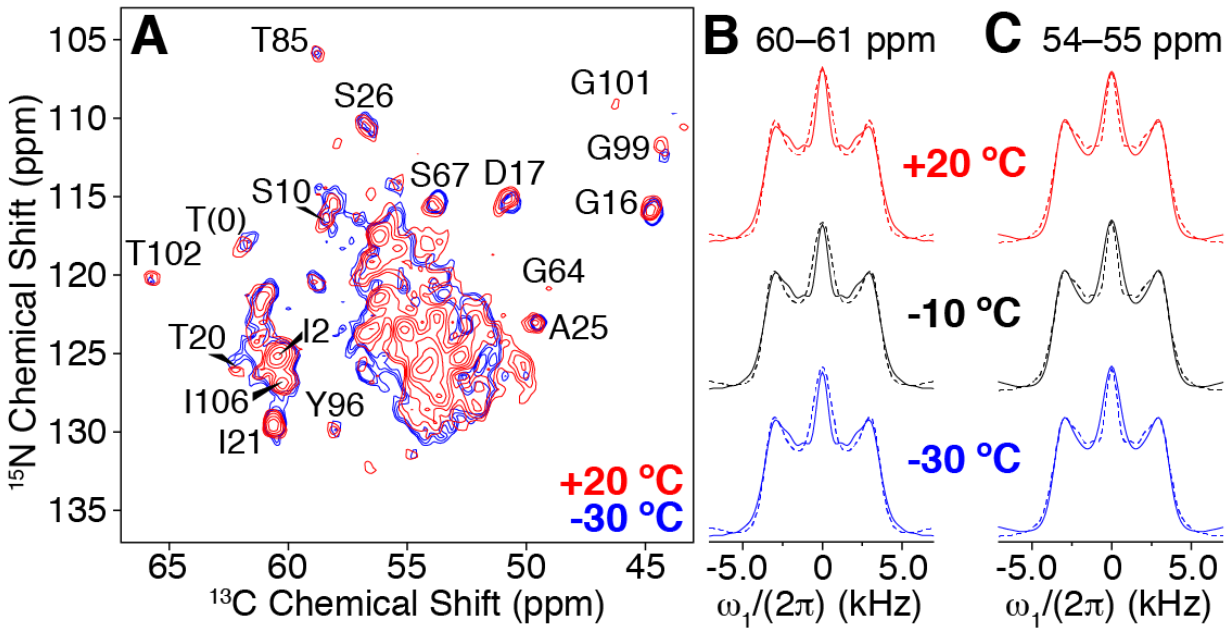

Figure 3.3. (A) Overlay of ^{15}N-^{13}Cα 2D spectra with short contact times collected at +20 ºC (red) and -30 ºC (blue). Individual spectra are shown in Figure 3.8. (B–C) ^{1}Hα-^{13}Cα dipolar coupling line shapes obtained from 2D T-MREV experiments at +20 ºC (red), -10 ºC (black), and -30 ºC (blue) are plotted along $\omega_1/(2\pi)$ from -5.0 to +5.0 kHz for integrated ^{13}Cα regions of (B) 60–61 ppm and (C) 54–55 ppm (experimental, solid; simulated, dashed). Fitting simulations for the shown line shapes all demonstrated order parameters of $S \geq 0.95$. Additional fittings are provided in Figure 3.9.

To extend upon these observations and obtain a quantitative measure of the protein dynamics, we performed ^{1}H-^{13}Cα dipolar coupling measurements for ^{13}Cα 1D regions between 50–63 ppm using the T-MREV[21,22] pulse sequence (Figures 3.3B–C and 3.9). The scaled dipolar couplings (and order parameters) corresponding to 1 ppm-wide integrals of the ^{13}Cα region were determined from fitting simulations (and normalization to NAV). Across the ^{13}Cα chemical shift range of 50 ppm to 63 ppm, the order parameters consistently demonstrated values of ≥ 0.95, supporting the presence of an extensive and highly ordered fibril. Furthermore, the order parameters for the ^{13}Cα spectral region exhibited no significant dependence on temperature (for all fitting simulations, see Figure 3.9). Together, these dipolar coupling data provide

strong support for the extensive rigidity of AL-09 V_L fibrils—encompassing the majority of the 108 aa protein sequence (~70 residues)[14]—as well as a low dependence on temperatures above and below the bulk water phase transition. It is interesting to note that this rather extensive structured region observed for AL-09 V_L fibrils is slightly larger than the sizes of several previously determined fibril structures, such as amyloid-β (21–40 residues), human prion protein variant huPrP23–144 (~28 residues), and α-synuclein fibrils (~56 residues), yet comparable in size to that observed for PI3-SH3 fibrils (75 residues) and the long straight fibrils of β_2-microglobulin (64 residues).[18,23-27]

3.4.2. The N- and C-terminal residues of AL-09 V_L are highly ordered in the fibril structure.

As demonstrated above, one prominent feature of AL-09 V_L fibrils is that a large proportion of the most highly ordered residues are located near the N- and C-termini of the protein sequence (Figures 3.2 and 3.7).[14] The ^{15}N-^{13}Cα-^{13}CX 3D strips in Figure 3.4A–B demonstrate the high resolution and sensitivity for a subset of the N- and C-terminal residues, supporting both their high rigidity and structural order. Furthermore, the observation of strong and well-resolved side-chain cross peaks for these residues provides substantial support for their adoption of a single, ordered structure in the fibril state, as the existence of many different conformations would cause side-chain signals arising from multiple magnetization transfer steps to be much weaker and broader. Additionally, these residues appear to exhibit a homogenous conformation, because only a single set of corresponding backbone and side-chain cross-peaks are observed in each 3D strip (Figure 3.4). For example, for residue V15, two unique chemical shifts (22.6 ppm and 20.3 ppm) are observed for the geminal side-chain methyl carbons (^{13}Cγ1 and ^{13}Cγ2), indicating that the residue only exists in one conformation (Figure 3.4A). Collectively, these observations of well-defined side-chain rotamers in the termini are particularly interesting in contrast to previous studies of other fibrils for which the termini are highly dynamic.[19,28] Given both the high rigidity and order of the N- and C-terminal residues, it is very likely that these two regions are major participants in maintaining the rigid structural integrity of the fibril state. These data further suggest that AL-09 V_L

fibrils demonstrate a long-range structural order which may be a unique feature of AL proteins relative to other reported amyloid fibril structures.[29,30]

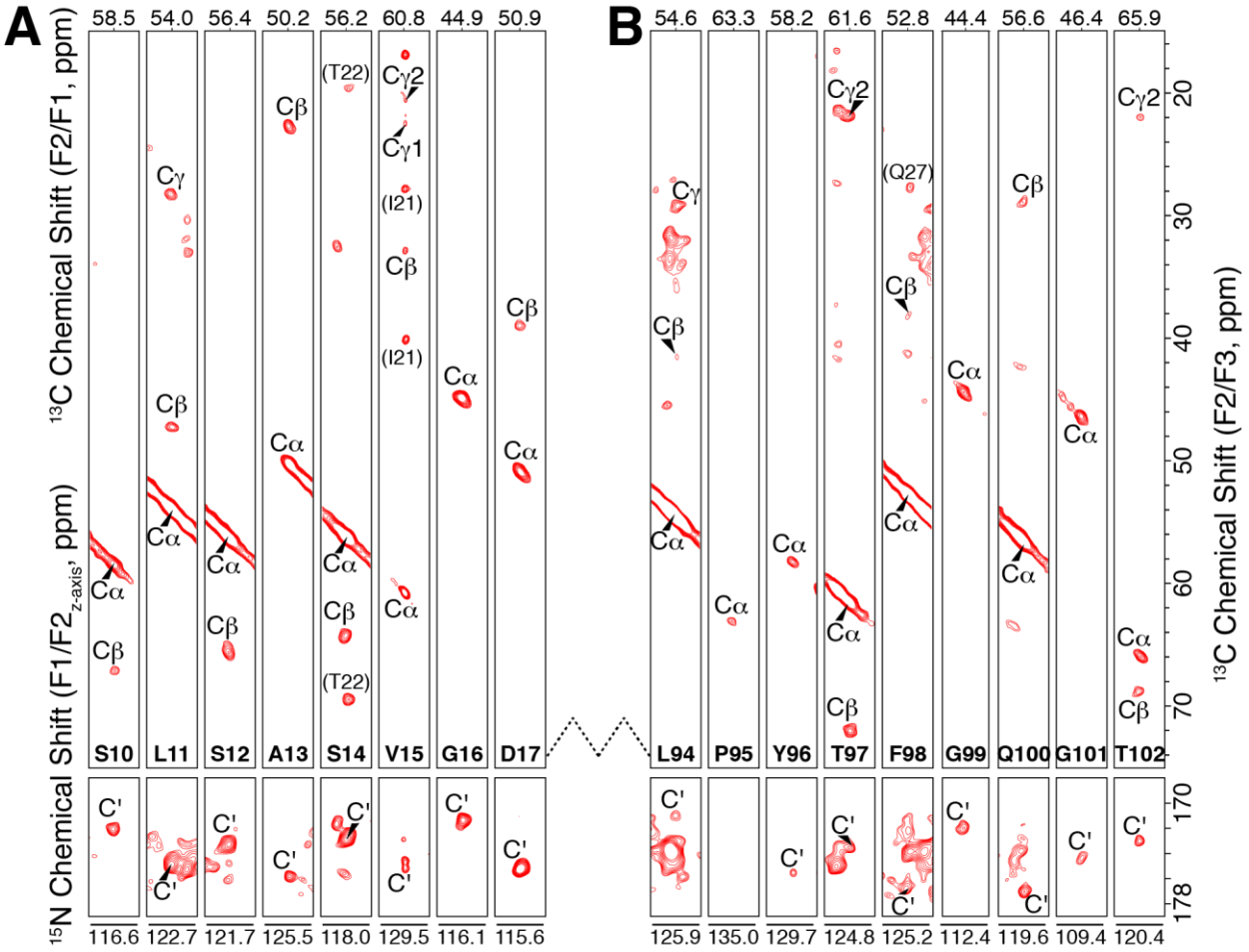

Figure 3.4. Strip plot from the 3D NCACX spectrum of [^{13}C,^{15}N]-labeled AL-09 V_L fibrils, demonstrating the backbone and side-chain atom assignments for (A) S10 to D17 (N-terminal residues) and (B) L94 to T102 (C-terminal residues). The sequential backbone atom connectivity of these and all other assigned residues is described previously.[14] Data were acquired at 600 MHz with a MAS rate of 13.333 kHz and DARR mixing time of 25 ms.

To quantify the rigidity of the termini at a site-specific level, we performed the 3D $R48_3$[18] version of R-symmetry sequences[31] with SPC-7 ^{13}C-^{13}C mixing[32] to measure the scaled ^{1}H-$^{13}C\alpha$ dipolar couplings of 14 total residues located at either the N- or C-terminus (resolved in a third dimension via the $^{13}C\beta$). Specifically, these residues include S7, S12–V15, V19–I21, and A25 (N-terminus) as well as A84, T85, L94, P95, and T97 (C-terminus), and were chosen because they exhibited strong $^{13}C\beta$-$^{13}C\alpha$ cross-peaks in the first CC 2D spectral plane (i.e., those above 15 times the noise floor) from which the dipolar dephasing trajectories could be readily extracted (Figure 3.5A). For all cases, the simulated ^{1}H-$^{13}C\alpha$ scaling factors corresponded to order parameters, $S \geq 0.95$ relative to the residue exhibiting the highest scaling factor, I21 (Figures 3.5B–C and 3.10). Collectively, the large magnitude of these backbone $^{13}C\alpha$ order parameters reveal the extensive and high rigidity of residues within both the N- and C-termini.

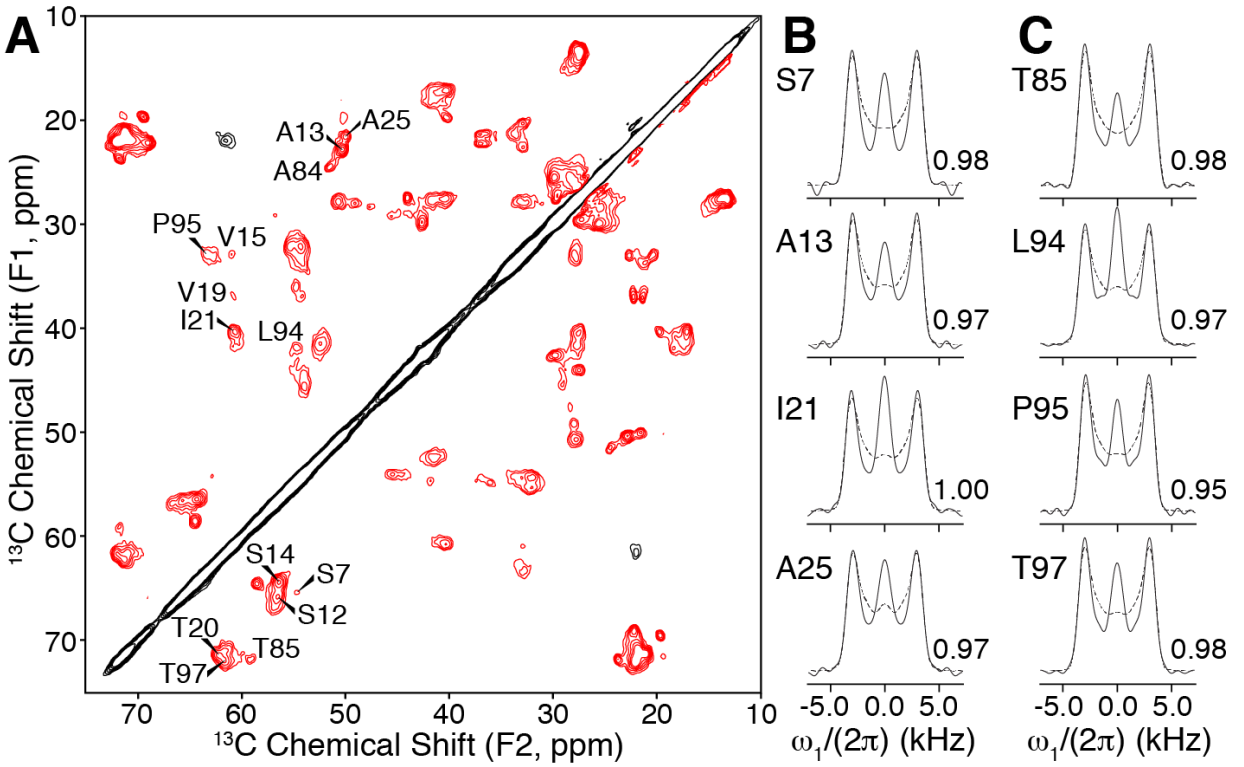

Figure 3.5. Site-specific ^{1}Hα-^{13}Cα dipolar coupling measurements for AL-09 V_L fibrils. (A) First ^{13}C-^{13}C 2D plane of the R48$_3^{18}$-symmetry 3D experiment, collected with SPC-7 q42 ^{13}C-^{13}C mixing at 20 ºC (black and red peaks indicate positive and negative signal intensity, respectively). Residues for which the ^{1}H-^{13}Cα dipolar coupling line shapes were extracted and simulated are labeled. (B–C) A subset of the experimental (solid) and simulated (dashed) line shapes illustrating the site-specific rigidity of several residues near the N- and C-termini (B and C, respectively) of the AL-09 V_L sequence. The corresponding order parameter, S, for each ^{13}Cα site is indicated next the line shape (normalized to the site with the largest measured ^{1}H-^{13}Cα scaling factor). Additional fitting simulations are shown in Figure 3.10.

Interestingly, as described previously on amyloidogenic/AL V_L domains, a "cryptic epitope" is proposed to occur within the first 18 residues of the N-terminus in which the presence of a proline at position 8 in combination with residues 1–4 and 13–18 supports the formation of an aberrant bend within β-strand A of the native V_L domain.[33] This conformational rearrangement of the N-terminus around P8

upon V_L misfolding would be consistent with our observation of its involvement in the fibril structure, as the CANCO CP efficiency generally increases in either direction for residues before and after P8 (Figure 3.7B). Furthermore, the presence of a β bulge in strand G at the C-terminus of V_L domains (residues 97–108) has been implicated in serving as a protective structural motif that discourages fibril formation,[34] with a particularly prominent β-bulge in kappa V_L domains over lambda structures.[35] It is possible that during fibril formation this motif undergoes a conformational change that supports an aberrant—yet highly ordered—conformation. Given that both the N- and C-termini of the AL-09 V_L sequence contain a ~12 aa stretch of generally intervening hydrophilic-hydrophobic residues (i.e., S10–T22 and Y96–K107), one entropically-favorable rearrangement that may occur could be through the interaction of these regions via their hydrophobic side chains. Certainly, understanding the specific orientation and interactions associated with the N- and C-terminal residues of AL-09 V_L fibrils is fundamental to determining the tertiary fibril structure, and may be pursued in future SSNMR studies by the measurement of intramolecular distance restraints[36,37] in combination with sparsely or selectively ^{13}C-labeled samples.[17,36,38-40]

Additional insight into the structure and formation of AL-09 V_L fibrils is provided by consideration of the effects of mutations on the protein stability. Globular proteins undergo partial unfolding in order to initiate misfolding events. This is unlike unstructured peptides (Aβ peptide) and intrinsically disordered proteins (α-synuclein) that initiate misfolding events from an unstructured or unfolded state. Mutations associated to amyloid diseases—in general—decrease the conformational stability of the globular native fold and promote aggregation *in vitro*. These mutations usually influence the thermodynamics of the process, and may not participate in direct interactions within the fibril structure.[41] Our mutational analyses of AL-09 V_L have shown that the non-conservative mutations located within the dimer interface of the soluble protein structure destabilize the protein and accelerate amyloid formation.[10] As non-conservative mutations in AL proteins may lower the energy barrier for the initial misfolding events to occur, we hypothesize that these effects are involved in the case of AL-09 V_L during

the amyloid nucleation event. Accordingly, in combination with our observation of possibly reduced CP signal intensity and/or resolution for peaks corresponding to mutation sites, this description suggests that the mutations have critical influence in the early misfolding event but may not be key contributors to maintaining the rigid final fibril structure. Such an explanation would be consistent with the extensive rigidity and order we report to primarily be localized in the N- and C-terminal regions of AL-09 V_L fibrils.

3.4.3. AL-09 V_L fibrils prepared *in vitro* exhibit structural similarities to AL extracts from human tissue.

In order to assess the physiological relevance of the AL-09 V_L fibril samples studied herein, we investigated the structural similarities to *in vivo* fibrils by comparing the (^{1}H)-^{13}C CP 1D spectrum of AL-09 V_L fibrils (U-[^{15}N,^{13}C]-labeled) with that of AL fibril deposits extracted directly from post-mortem human spleen tissue (HIG; natural abundance) (Figure 3.6A). These *ex vivo* fibrils share the same $\varkappa$I germline gene product with AL-09 V_L, which minimizes the number of amino acid differences between the two samples (Figure 3.6B). As illustrated in Figure 3.6A, the HIG fibrils are well structured and exhibit similarities in chemical shift patterns as AL-09 V_L fibrils, such as with respect to the downfield Thr- and Ser-^{13}Cβ chemical shifts indicative of predominantly a β-sheet secondary structure. Furthermore, unique glycosaminoglycan (GAG) signals are observed near 75 ppm and 183 ppm in the HIG spectrum (arising from the hydroxyl-containing and carboxylate carbons of GAGs, respectively), consistent with mass spectrometric analysis which revealed the presence of heparan sulfate proteoglycan (HSPG2) in the spleen tissue extract. Together, these results are in agreement with the known association of glycosaminoglycans with AL fibrils *in vivo*.[42,43] Other spectral differences can be accounted for by variations in protein sequence of AL-09 V_L and HIG (Figure 3.6B).

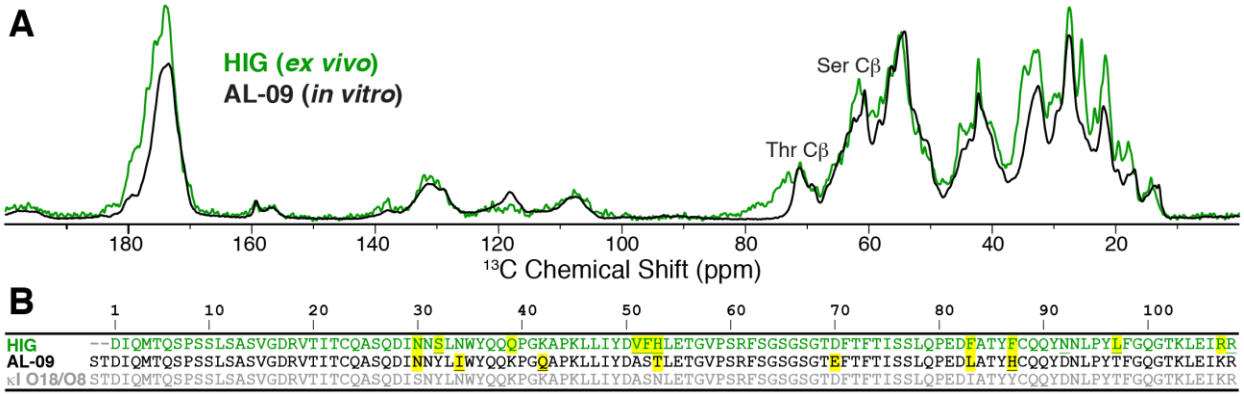

Figure 3.6. Comparison of AL-09 V_L fibrils with AL spleen tissue extract (HIG). (A) Overlay of ^{13}C 1D SSNMR spectra (normalized to $^{13}C\alpha$ region) for ^{13}C natural abundance amyloid extract from postmortem splenic tissue (HIG; green) and AL-09 V_L (black) fibrils. (B) Sequence alignment for HIG, AL-09, and κI O18/O8 (germline; gray) V_L proteins (residues –1 and 0 are purification artifacts). Highlighted residues indicate mutation sites; those that are non-conservative are underlined.

In addition to providing foundational insight on the structural similarities of two AL fibrils, these results represent the first demonstration of the feasibility of studying the molecular structures of AL fibrils that are pathologically relevant. Moreover, although the spectral sensitivity of *ex vivo* fibrils is limited by the lack of ^{13}C and ^{15}N enrichment, future structural studies of tissue fibril deposits will be facilitated by using the fibril material as a seed for *in vitro* preparations of isotopically-labeled fibril samples. This approach has been shown to be successful in previous SSNMR studies for $A\beta_{1-40}$ fibrils extracted from Alzheimer's disease afflicted brain tissue.[44] These studies will be further supported by the ability to immediately assess the physiological relevance and structural similarity of the labeled fibrils using the (^1H)-^{13}C CP 1D spectrum obtained for HIG fibrils here as "fingerprint" comparison (Figure 3.6A).

As described earlier, one inherent challenge associated with the study of AL disease is the large sequence variation that can occur between patients, consequently limiting the ability to predict and identify early-on the occurrence of light-chain sequences with high fibril-forming propensity.

Interestingly, however, previous work has identified a murine monoclonal antibody that can be used to specifically bind light chain fibrils (as opposed to the native soluble species) irrespective of slight sequence variations—even between germline isotypes—so long as a proline is present at position 8 of the protein sequence.[33,45] These observations indicate that a unique structural motif forms around P8—previously proposed to be a β-turn—that binds specifically to the binding domain of the murine antibody. Given that P8 is 100% conserved among kappa light-chain sequences such as AL-09 and HIG (Figure 3.6B) and that it exhibits structural order in AL-09 V_L fibrils, it is possible that the corresponding structure of the N-terminal region represents this unique recognition motif. While additional structural studies will be necessary to confirm this, the agreement of these two structures would suggest that all AL fibrils containing this proline-8 would adopt the same structural motif and possibly other similar aspects of the structural framework of the fibril. Certainly, knowledge of whether a prevailing structural conformation or motif exists between different AL fibrils would provide enormous support for the structure-based design of potential diagnostic or therapeutic strategies, thus further emphasizing the importance of future structural studies on pathologically relevant AL fibrils.

3.5. CONCLUSIONS

In this study, we investigated the secondary structure and relative dynamics across the AL-09 V_L sequence in its fibril form through the use of MAS SSNMR spectroscopy, and subsequently assessed the structural similarity of *ex vivo* AL fibrils obtained from spleen tissue with *in vitro*-generated AL-09 V_L fibrils used in the study here. Overall, our data reveal that extensive regions of the AL-09 V_L sequence are highly rigid and β-strand rich in the fibril state, encompassing the majority of the residues in the protein. Further, our results indicate that the residues near the N- and C-termini are especially well-ordered and exhibit only a single conformation in the fibril form. These measurements of fibril rigidity and relative order were performed at both the global sample and site-specific levels. In light of the success of previous work on the determination of other fibril structures, such as amyloid-β in Alzheimer's disease and α-synuclein in Parkinson's disease,[23,25] the results presented here on AL-09 V_L fibrils are very promising

with respect to the feasibility of determining the corresponding three-dimensional structure. Furthermore, the unique structural features already observed for AL-09 V_L fibrils—including the high degree of order for N- and C-terminal residues—relative to other fibril structures suggests that AL-09 exhibits a different and possibly novel type of fold or packing. Moreover, the serine-rich sequence of AL-09 V_L relative to the generally hydrophobic-rich sequences of other amyloid fibrils[46,47] may lead to the contribution of additional complexities to the fibril structure.

In comparing the one-dimensional ^{13}C spectrum of *in* vitro-prepared AL-09 V_L fibrils with that of *ex vivo* AL fibrils obtained from human spleen tissue, our results demonstrate that both fibrils exhibit a high degree of rigidity and β-strand content. Importantly, these data also set the framework on which additional pathologically relevant AL samples may be studied, including the *in vitro* preparation of isotopically-labeled AL fibrils by the introduction of a fibril seed from the tissue extract to propagate the relevant structure. Indeed, given the recent identification of AL fibrils as the toxic species to cardiomyocytes *in vitro* by causing cell growth arrest,[5] there is tremendous value in the delineation of the molecular fibril structure for benefiting the pursuit of AL therapies that are directed at the fibril species. Accordingly, we envision that these structural insights into AL fibrils will inform the ongoing search for effective treatments and biomarkers of AL disease, including small molecule inhibitors of fibril formation as well as monoclonal antibodies specific for insoluble light chain fibrils.[4,8,9,33,48]

3.6. EXPERIMENTAL SECTION

Protein Expression and Purification. Uniformly [^{13}C,^{15}N]-labeled AL-09 V_L protein (a light-chain variable domain derived from a cardiac AL patient)[2] was expressed and purified utilizing a modification of earlier methods[4,49] to accommodate minimal media with ^{13}C and ^{15}N isotopes, as described.[14] *Escherichia coli* BL21 (DE3) Gold competent cells (Stratagene, La Jolla, CA), were used to express the AL-09 V_L plasmid (pET12a) in 2xYT media (Sigma-Aldrich, Saint Louis, MO), and grown until an A_{600nm} of 0.6. The culture was subsequently transferred into minimal M9 media[50] (with labeled [^{13}C]-glucose and [^{15}N]-ammonium chloride; 3.0 g and 1.0 g per L media, respectively) at a four-fold higher

cell concentration, and incubated for 1 h at 37 °C before induction with IPTG (0.8 mM). Expression was allowed to persist overnight at 25 °C and 100 rpm prior to harvesting. Protein purification was performed using osmotic shock (as described),[14] followed by injection onto a HiLoad 16/60 Superdex 75 size-exclusion column in 10 mM Tris–HCl, pH 7.4 buffer using an AKTA FPLC system (GE Healthcare; Piscataway, NJ).

AL-09 V_L purity and concentration were assessed by SDS-PAGE and UV absorption (extinction coefficient, $\varepsilon = 13{,}610\ cm^{-1}\ M^{-1}$), respectively. Far UV circular dichroism (CD) spectroscopy was used to verify native protein folding prior to the fibril formation reaction, acquiring from 200–260 nm at 4 °C with a Jasco Spectropolarimeter 810 (JASCO, Inc., Easton, MD) and a 0.2 cm path-length quartz cuvette. Percent [^{13}C, ^{15}N]-isotope incorporation in purified AL-09 V_L was determined by MALDI-TOF to be approximately 99%.

***In vitro* Fibril Formation.** Uniformly [^{13}C, ^{15}N]-labeled AL-09 V_L fibril samples were formed from purified protein monomer as described.[14,51] AL-09 V_L monomer was separated from any pre-formed aggregates by filtering the purified product through a 0.45 μm membrane and subjecting the filtrate to ultracentrifugation (90,000 rpm, 3.3 h) at 4 ºC using a NVT-90 rotor and Optima L-100 XP centrifuge (Beckman Coulter, Brea, CA). Fibril-forming samples were subsequently prepared on ice from the monomer-containing supernatant at 20 μM in Acetate Borate Citrate buffer (10 mM each, pH 2.0), containing 150 mM NaCl and 0.02% NaN_3 in 1.5 mL low-binding microcentrifuge tubes (1.0 mL total volume). An additional sample was prepared for each replicate under the same conditions, but with 10 μM Thioflavin T (ThT) added for monitoring fibrillation progress. Fibril formation reactions were performed in triplicate in sealed (Nunc, Roskilde, Denmark), black 96-well polystyrene plates (Greiner, Monroe, NC) at 37 °C and 300 rpm (New Brunswick Scientific Innova40 incubator shaker), using 260 μL aliquots per well until fibril formation was complete based on the change in enhanced fluorescence intensity of ThT. The presence of fibrils was subsequently confirmed by transmission electron

microscopy (TEM), according to previous protocols.[14] The average time for AL-09 V_L fibrils to form was approximately two days.

Preparation of Fibrils for SSNMR Studies. As previously described,[14] completed fibril reaction mixtures were subjected to ultracentrifugation in polypropylene microfuge tubes using a table-top ultracentrifuge with a TLA-100.3 (Beckman Coulter) rotor and microfuge tube adapters for 1 h at 4 ºC and 55,000 rpm (~100,000 *g*). Pelleted fibril material was separated from the supernatant and washed twice with deionized water using a hand-homogenizer and repeated ultracentrifugation. Washed fibril pellets were dried completely under nitrogen gas until the sample mass remained unchanged. Dried AL-09 V_L fibril samples were packed into either a 3.2 mm (outer-diameter) standard-wall or thin-wall SSNMR MAS rotor (Agilent Technologies, Santa Clara, CA) and hydrated to a level of 50% water by mass. Hydration of packed samples was retained using Kel-F and rubber spacers on both ends of the sample in the rotor.

AL Amyloid Extraction. Patient HIG was a male Caucasian diagnosed with multiple myeloma and subsequently hepatosplenic ALκ amyloidosis. The light chain component comprising the amyloid fibrils was shown by amino acid (aa) sequencing to be a κI protein expressed by the O18-O8 (IGKV-1-33) germline gene with a glycosylation site introduced at N30-N31-S32.

Water-soluble amyloid extracts were prepared following prior protocols with some modifications.[45,52] In short, 30–40 g of freshly-frozen (-80 °C) spleen obtained postmortem from patient HIG were homogenized in about 300 ml pre-chilled saline solution (Virtis-Tempest; Virtis, Gardiner, NY). The resulting homogenate was centrifuged for 30 min at 6 °C and 15,000 *g* and residual saline-soluble material was removed by repeated cycles of homogenization and washing until the optical density (at 280 nm) of the supernatant was less than 0.10. The pellet was then subjected to repeated rounds of homogenization, washing (cold, de-ionized water), and centrifugation, collecting and lyophilizing the amyloid-containing supernatant after each successive round. The amount of protein recovered represented ~25% the weight of the starting material. Approximately 10 mg of the lyophilized material were

rehydrated to ~50% water by mass and packed into a 3.2 mm standard-wall SSNMR rotor (Agilent Technologies).

Solid-State NMR Spectroscopy. MAS SSNMR data of [^{13}C,^{15}N]-labeled AL-09 V_L were collected on the following spectrometers: a 600 MHz InfinityPlus wide-bore or narrow-bore spectrometer equipped with a 3.2 mm Balun ^{1}H-^{13}C-^{15}N or T3 ^{1}H-^{13}C-^{31}P probe, respectively; a 500 MHz VNMRS (Varian, Walnut Creek, CA) spectrometer equipped with a 3.2 mm Balun ^{1}H-^{13}C-^{15}N probe; and a 750 MHz VNMRS spectrometer equipped with a 3.2 mm Balun or BioMAS ^{1}H-^{13}C-^{15}N probe (Varian, Fort Collins, CO). Multidimensional ^{15}N and ^{13}C correlation experiments were acquired using a MAS rate of 11.111 kHz (500 MHz spectrometer), 13.333 kHz (600 MHz), or 12.500 kHz (750 MHz), and a variable-temperature (VT) stack air flow set-point of -10 ºC, 0 ºC or 10 ºC (for sample temperatures of -5 ºC, 4 ºC or 12 ºC ± 3 ºC, respectively, as determined by calibration with ethylene glycol).[53] A SPINAL-64 ^{1}H-decoupling power of 65–80 kHz was applied during evolution and acquisition periods.[54] One-bond ^{13}C-^{13}C correlations were achieved using either dipolar-assisted rotational resonance (DARR; τ_{mix} = 25–90 ms)[55] or supercycled POST-C7 (SPC-7; τ_{mix} = 0.9–1.8 ms)[32] for magnetization transfer, and one-bond ^{15}N-^{13}C or ^{13}C-^{15}N correlations were accomplished using SPECIFIC-CP.[56] Additional experimental details for multidimensional data acquisition (e.g., digitization settings and specific radio-frequency powers) are summarized previously.[14]

Temperature dependent ^{15}N-^{13}Cα (NCA) 2D experiments were performed at -30 ºC, -10 ºC, and 20 ºC (VT set points, corresponding to sample temperatures of -23 ºC, -5 ºC, and 21 ºC ± 3 ºC, respectively), using a MAS rate of 13.333 kHz (600 MHz) and SPINAL-64 ^{1}H-decoupling power[13] of 72 kHz. Short cross-polarization contact times of 0.1 ms and 2.0 ms were used for ^{1}H-^{15}N and ^{15}N-^{13}Cα transfers, respectively, to promote only intraresidue magnetization transfer (relative to other ^{1}H-^{15}N-^{13}C correlation experiments which used times of 1.5 ms and 5.0 ms, respectively). T-MREV ^{1}H-^{13}C dipolar recoupling experiments[22] were performed at the same three temperatures (VT), but using a MAS rate of 7.576 kHz (600 MHz) and TPPM ^{1}H decoupling power[18] of 72 kHz. Scaling factors for ^{1}H-^{13}C dipolar couplings of AL-09 V_L fibrils were normalized to the ^{1}H–^{13}Cα scaling factor measured for crystalline [^{13}C, ^{15}N]-

labeled N-acetyl-valine. $R48_3^{18}$-symmetry ^{1}H-^{13}C-^{13}C-^{1}H 3D dipolar recoupling experiments were performed using a MAS rate of 13.333 kHz (600 MHz) and VT set-point of 20 ºC, with a SPINAL-64 ^{1}H-decoupling power of 74 kHz. MAS SSNMR data for the spleen tissue AL extract sample (HIG) were collected on the 600 MHz narrow-bore spectrometer using a MAS rate of 10.000 kHz, a VT set-point of 10 ºC, and a SPINAL-64 ^{1}H-decoupling power of 65 kHz. (^{1}H)-^{13}C CP 1D spectrum represents a total of 61.5 h of data collection time (108,544 scans).

Solid-State NMR Data Processing and Analysis. SSNMR data were processed using NMRPipe.[57] Multidimensional ^{15}N and/or ^{13}C correlation spectra were analyzed using Sparky,[58] and chemical shift-based predictions on protein secondary structure were performed using TALOS-N[15] as well as comparison to the random coil chemical shift values presented by Zhang et al.[16] T-MREV and $R48_3^{18}$-symmetry ^{1}H-^{13}C dipolar coupling dephasing curves were analyzed using an in-house fitting program (written in FORTRAN).[22] The zero-frequency intensity present in R-symmetry spectra was excluded from the fitting simulations.[59]

3.7. ASSOCIATED CONTENT

3.7.1. **Accession Numbers**

Backbone and side-chain (partial) ^{13}C and ^{15}N chemical shift assignments for AL-09 V_L fibrils are available on the BioMagResBank (BMRB) under accession number 26879. Information on the light chain cDNA of AL-09 V_L is available on GenBank under accession number AF490909.

3.7.2. **Supporting Information**

TALOS-N predicted backbone dihedral angles and relative CANCO peak heights (Figure 3.7), the un-overlaid set of ^{15}N-$^{13}C\alpha$ 2D spectra at -30 ºC, -10 ºC, and 20 ºC (Figure 3.8), and the complete set of ^{1}H-$^{13}C\alpha$ T-MREV (Figure 3.9) and $R48_3^{18}$-symmetry (Figure 3.10) dipolar coupling line shapes.

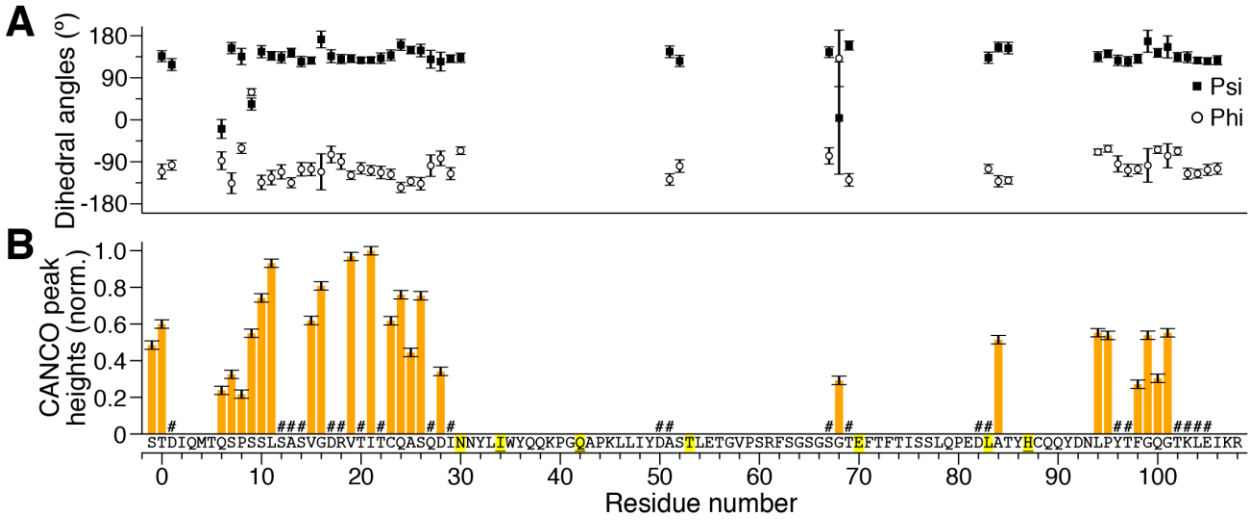

Figure 3.7. (A) TALOS-N predicted backbone dihedral angles for AL-09 V_L fibrils, plotted as a function of residue number (psi, closed squares; phi, open circles); error bars indicate the standard deviation of the averaged TALOS-N matches. (B) Normalized peak heights from the cross-peaks of the CANCO 3D spectrum of U-[^{13}C,^{15}N]-labeled AL-09 fibrils. Residue number corresponds to that of the carbonyl resonance; error bars correspond to the estimated spectral noise. Residues labeled with a # symbol were excluded due to partial peak overlap in the CANCO 3D spectrum. Highlighted residues represent mutation sites; those that are underlined are non-conservative.

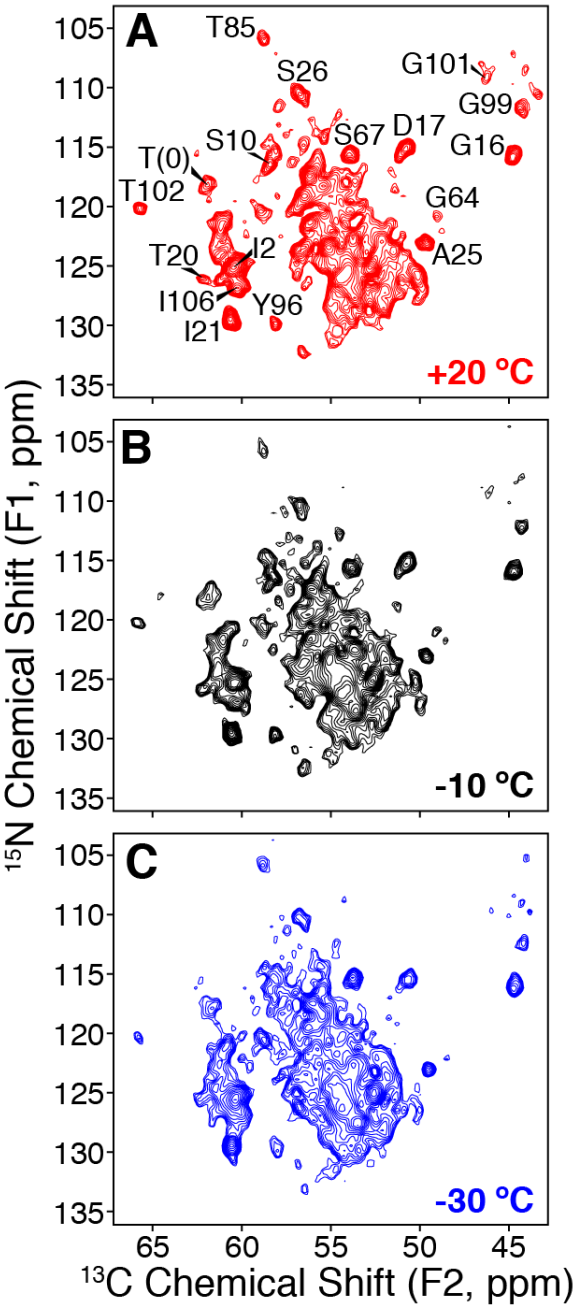

Figure 3.8. ^{15}N-^{13}Cα 2D spectra of AL-09 V_L fibrils with short contact times collected at +20 ºC (A; red), -10 ºC (B; black), and -30 ºC (C; blue). Spectral contours are cut at seven times the noise floor.

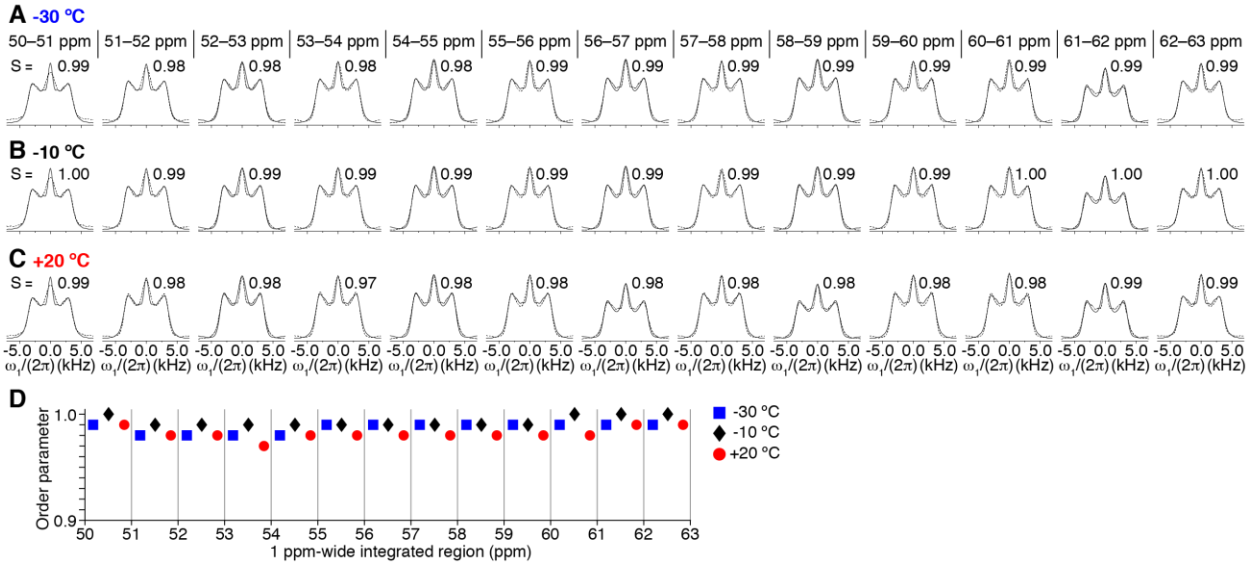

Figure 3.9. ^{1}Hα-^{13}Cα T-MREV experimental (solid) and simulated (dashed) fittings for a range of 1 ppm-wide ^{13}Cα regions (between 50–63 ppm) for AL-09 V_L fibrils at (A) -30 ºC, (B) -10 ºC, and (C) +20 ºC. Dipolar coupling line shapes are plotted along $\omega_1/(2\pi)$ from -5.0 to +5.0 kHz. Data were collected at 600 MHz (^{1}H) and a MAS rate of 7,576 Hz, and all line shapes were fit assuming a ^{13}Cα spin-topology coupled to three protons (^{15}N(H)-^{13}Cα(H)-^{13}Cβ(H)$_x$). (D) ^{1}Hα-^{13}Cα dipolar order parameters for T-MREV fitting simulations of 1 ppm-wide integrated ^{13}Cα regions between 50 ppm to 63 ppm, at -30 ºC (blue squares), -10 ºC (black diamonds), and +20 ºC (red circles).

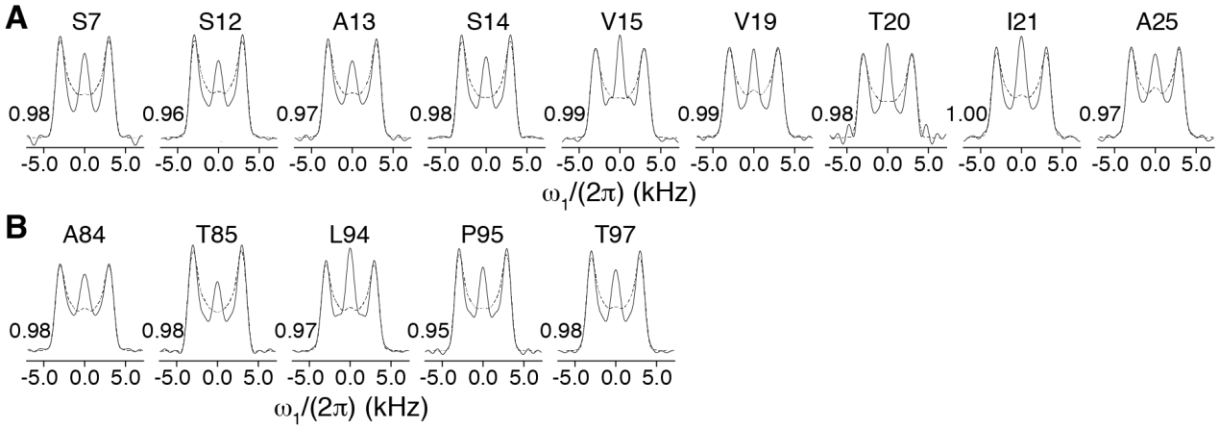

Figure 3.10. Site-specific ^{1}H-(^{13}Cβ/')-^{13}Cα-^{1}H dipolar coupling line shapes (experimental, solid; simulated fit, dashed) obtained from ^{1}H-^{13}C-^{13}C-^{1}H R48$_3^{18}$-symmetry 3D (SPC-7 q42 ^{13}C-^{13}C mixing) data for AL-09 V_L residues located near the (A) N-terminus and (B) C-terminus. The corresponding order parameter, S, for each ^{13}Cα site is indicated next the line shape, and is normalized to the residue exhibiting the highest ^{1}Hα-^{13}Cα scaling factor (i.e., I21).

3.8. REFERENCES

(1) Baden, E. M.; Sikkink, L. A.; Ramirez-Alvarado, M. Light chain amyloidosis – current findings and future prospects. *Curr. Protein Pept. Sci.* **2009**, *10*, 500–508.

(2) Abraham, R. S.; Geyer, S. M.; Price-Troska, T. L.; Allmer, C.; Kyle, R. A.; Gertz, M. A.; Fonseca, R. Immunoglobulin light chain variable (V) region genes influence clinical presentation and outcome in light chain-associated amyloidosis (AL). *Blood* **2003**, *101*, 3801–3808.

(3) Levinson, R. T.; Olatoye, O. O.; Randles, E. G.; Howell, K. G.; DiCostanzo, A. C.; Ramirez-Alvarado, M. Role of mutations in the cellular internalization of amyloidogenic light chains into cardiomyocytes. *Sci. Rep.* **2013**, *3*, 1–8.

(4) McLaughlin, R. W.; De Stigter, J. K.; Sikkink, L. A.; Baden, E. M.; Ramirez-Alvarado, M. The effects of sodium sulfate, glycosaminoglycans, and Congo red on the structure, stability, and amyloid formation of an immunoglobulin light-chain protein. *Protein Sci.* **2006**, *15*, 1710–1722.

(5) Marin-Argany, M.; Lin, Y.; Misra, P.; Williams, A.; Wall, J. S.; Howell, K. G.; Elsbernd, L. R.; McClure, M.; Ramirez-Alvarado, M. Cell Damage in Light Chain Amyloidosis: Fibril Internalization, Toxicity, and Cell-Mediated Seeding. *J. Biol. Chem.* **2016**, *291*, 19813–19825.

(6) McWilliams-Koeppen, H. P.; Foster, J. S.; Hackenbrack, N.; Ramirez-Alvarado, M.; Donohoe, D.; Williams, A.; Macy, S.; Wooliver, C.; Wortham, D.; Morrell-Falvey, J.; Foster, C. M.; Kennel, S. J.; Wall, J. S. Light Chain Amyloid Fibrils Cause Metabolic Dysfunction in Human Cardiomyocytes. *PLoS One* **2015**, *10*, e0137716.

(7) Cooley, C. B.; Ryno, L. M.; Plate, L.; Morgan, G. J.; Hulleman, J. D.; Kelly, J. W.; Wiseman, R. L. Unfolded protein response activation reduces secretion and extracellular aggregation of amyloidogenic immunoglobulin light chain. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 13046–13051.

(8) Gertz, M. A.; Landau, H.; Comenzo, R. L.; Seldin, D.; Weiss, B.; Zonder, J.; Merlini, G.; Schonland, S.; Walling, J.; Kinney, G. G.; Koller, M.; Schenk, D. B.; Guthrie, S. D.; Liedtke, M. First-in-Human Phase I/II Study of NEOD001 in Patients With Light Chain Amyloidosis and Persistent Organ Dysfunction. *J. Clin. Oncol.* **2016**, *34*, 1097–1103.

(9) Gertz, M. A.; Landau, H. J.; Weiss, B. M. Organ response in patients with AL amyloidosis treated with NEOD001, an amyloid-directed monoclonal antibody. *Am. J. Hematol.* **2016**, *91*, E506–E508.

(10) Baden, E. M.; Randles, E. G.; Aboagye, A. K.; Thompson, J. R.; Ramirez-Alvarado, M. Structural insights into the role of mutations in amyloidogenesis. *J. Biol. Chem.* **2008**, *283*, 30950–30956.

(11) Baden, E. M.; Owen, B. A.; Peterson, F. C.; Volkman, B. F.; Ramirez-Alvarado, M.; Thompson, J. R. Altered dimer interface decreases stability in an amyloidogenic protein. *J. Biol. Chem.* **2008**, *283*, 15853–15860.

(12) Mukherjee, S.; Pondaven, S. P.; Jaroniec, C. P. Conformational Flexibility of a Human Immunoglobulin Light Chain Variable Domain by Relaxation Dispersion Nuclear Magnetic Resonance Spectroscopy: Implications for Protein Misfolding and Amyloid Assembly. *Biochemistry* **2011**, *50*, 5845–5857.

(13) Peterson, F. C.; Baden, E. M.; Owen, B. A.; Volkman, B. F.; Ramirez-Alvarado, M. A single mutation promotes amyloidogenicity through a highly promiscuous dimer interface. *Structure* **2010**, *18*, 563–570.

(14) Piehl, D. W.; Blancas-Mejía, L. M.; Ramirez-Alvarado, M.; Rienstra, C. M. Solid-state NMR chemical shift assignments for AL-09 VL immunoglobulin light chain fibrils. *Biomol. NMR Assign.* [Online Early Access]. DOI: 10.1007/s12104-016-9718-3. Published Online: 2016.

(15) Shen, Y.; Bax, A. Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J. Biomol. NMR* **2013**, *56*, 227–241.

(16) Zhang, H. Y.; Neal, S.; Wishart, D. S. RefDB: A database of uniformly referenced protein chemical shifts. *J. Biomol. NMR* **2003**, *25*, 173–195.

(17) Comellas, G.; Rienstra, C. M. Protein structure determination by magic-angle spinning solid-state NMR, and insights into the formation, structure, and stability of amyloid fibrils. *Annu. Rev. Biophys.* **2013**, *42*, 515–536.

(18) Helmus, J. J.; Surewicz, K.; Nadaud, P. S.; Surewicz, W. K.; Jaroniec, C. P. Molecular conformation and dynamics of the Y145Stop variant of human prion protein. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 6284–6289.

(19) Comellas, G.; Lemkau, L. R.; Nieuwkoop, A. J.; Kloepper, K. D.; Ladror, D. T.; Ebisu, R.; Woods, W. S.; Lipton, A. S.; George, J. M.; Rienstra, C. M. Structured regions of α-synuclein fibrils Include the early-onset Parkinson's disease mutation sites. *J. Mol. Biol.* **2011**, *411*, 881–895.

(20) Kloepper, K. D.; Zhou, D. H.; Li, Y.; Winter, K. A.; George, J. M.; Rienstra, C. M. Temperature-dependent sensitivity enhancement of solid-state NMR spectra of α-synuclein fibrils. *J. Biomol. NMR* **2007**, *39*, 197–211.

(21) Hohwy, M.; Jaroniec, C. P.; Reif, B.; Rienstra, C. M.; Griffin, R. G. Local structure and relaxation in solid-state NMR: accurate measurement of amide N-H bond lengths and H-N-H bond angles. *J. Am. Chem. Soc.* **2000**, *122*, 3218–3219.

(22) Rienstra, C. M.; Hohwy, M.; Mueller, L. J.; Jaroniec, C. P.; Reif, B.; Griffin, R. G. Determination of multiple torsion-angle constraints in U–13C,15N-labeled peptides: 3D 1H–15N–13C–1H dipolar chemical shift NMR spectroscopy in rotating solids. *J. Am. Chem. Soc.* **2002**, *124*, 11908–11922.

(23) Petkova, A. T.; Ishii, Y.; Balbach, J. J.; Antzutkin, O. N.; Leapman, R. D.; Delaglio, F.; Tycko, R. A structural model for Alzheimer's β-amyloid fibrils based on experimental constraints from solid state NMR. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 16742–16747.

(24) Tycko, R. Physical and structural basis for polymorphism in amyloid fibrils. *Protein Sci.* **2014**, *23*, 1528–1539.

(25) Tuttle, M. D.; Comellas, G.; Nieuwkoop, A. J.; Covell, D. J.; Berthold, D. A.; Kloepper, K. D.; Courtney, J. M.; Kim, J. K.; Barclay, A. M.; Kendall, A.; Wan, W.; Stubbs, G.; Schwieters, C. D.; Lee, V. M. Y.; George, J. M.; Rienstra, C. M. Solid-state NMR structure of a pathogenic fibril of full-length human α-synuclein. *Nat. Struct. Mol. Biol.* **2016**, *23*, 409–415.

(26) Bayro, M. J.; Maly, T.; Birkett, N. R.; Macphee, C. E.; Dobson, C. M.; Griffin, R. G. High-resolution MAS NMR analysis of PI3-SH3 amyloid fibrils: backbone conformation and implications for protofilament assembly and structure. *Biochemistry* **2010**, *49*, 7474–7484.

(27) Debelouchina, G. T.; Platt, G. W.; Bayro, M. J.; Radford, S. E.; Griffin, R. G. Magic angle spinning NMR analysis of β2-microglobulin amyloid fibrils in two distinct morphologies. *J. Am. Chem. Soc.* **2010**, *132*, 10414–10423.

(28) Su, Y. C.; Sarell, C. J.; Eddy, M. T.; Debelouchina, G. T.; Andreas, L. B.; Pashley, C. L.; Radford, S. E.; Griffin, R. G. Secondary structure in the core of amyloid fibrils formed from human β(2)m and its truncated variant ΔN6. *J. Am. Chem. Soc.* **2014**, *136*, 6313–6325.

(29) Tjernberg, L. O.; Callaway, D. J. E.; Tjernberg, A.; Hahne, S.; Lilliehook, C.; Terenius, L.; Thyberg, J.; Nordstedt, C. A molecular model of Alzheimer amyloid β-peptide fibril formation. *J. Biol. Chem.* **1999**, *274*, 12619–12625.

(30) Rodriguez, J. A.; Ivanova, M. I.; Sawaya, M. R.; Cascio, D.; Reyes, F. E.; Shi, D.; Sangwan, S.; Guenther, E. L.; Johnson, L. M.; Zhang, M.; Jiang, L.; Arbing, M. A.; Nannenga, B. L.; Hattne, J.; Whitelegge, J.; Brewster, A. S.; Messerschmidt, M.; Boutet, B.; Sauter, N. K.; Gonen, T.; Eisenberg, D. S. Structure of the toxic core of α-synuclein from invisible crystals. *Nature* **2015**, *525*, 486–490.

(31) Zhao, X.; Eden, M.; Levitt, M. H. Recoupling of heteronuclear dipolar interactions in solid-state NMR using symmetry-based pulse sequences. *Chem. Phys. Lett.* **2001**, *342*, 353–361.

(32) Hohwy, M.; Rienstra, C. M.; Jaroniec, C. P.; Griffin, R. G. Fivefold symmetric homonuclear dipolar recoupling in rotating solids: Application to double quantum spectroscopy. *J. Chem. Phys.* **1999**, *110*, 7983–7992.

(33) O'Nuallain, B.; Allen, A.; Kennel, S. J.; Weiss, D. T.; Solomon, A.; Wall, J. S. Localization of a conformational epitope common to non-native and fibrillar immunoglobulin light chains. *Biochemistry* **2007**, *46*, 1240–1247.

(34) Hernandez-Santoyo, A.; Yauner, L. D.; Fuentes-Silva, D.; Ortiz, E.; Rudino-Pinera, E.; Sanchez-Lopez, R.; Horjales, E.; Becerril, B.; Rodriguez-Romero, A. A single mutation at the sheet switch region results in conformational changes favoring λ6 light-chain fibrillogenesis. *J. Mol. Biol.* **2010**, *396*, 280–292.

(35) Blancas-Mejía, L. M.; Ramirez-Alvarado, M. Systemic amyloidoses. *Annu. Rev. Biochem.* **2013**, *82*, 745–774.

(36) De Paepe, G.; Lewandowski, J. R.; Loquet, A.; Bockmann, A.; Griffin, R. G. Proton assisted recoupling and protein structure determination. *J. Chem. Phys.* **2008**, *129*, 245101.

(37) Gullion, T.; Schaefer, J. Rotational-echo double-resonance NMR. *J. Magn. Reson.* **1989**, *81*, 196–200.

(38) Higman, V. A.; Flinders, J.; Hiller, M.; Jehle, S.; Markovic, S.; Fiedler, S.; van Rossum, B. J.; Oschkinat, H. Assigning large proteins in the solid state: a MAS NMR resonance assignment strategy using selectively and extensively 13C-labelled proteins. *J. Biomol. NMR* **2009**, *44*, 245–260.

(39) Hong, M.; Jakes, K. Selective and extensive 13C labeling of a membrane protein for solid-state NMR investigations. *J. Biomol. NMR* **1999**, *14*, 71–74.

(40) Hiller, M.; Higman, V. A.; Jehle, S.; van Rossum, B. J.; Kuhlbrandt, W.; Oschkinat, H. [2,3-C-13]-labeling of aromatic residues-getting a head start in the magic-angle-spinning NMR assignment of membrane proteins. *J. Am. Chem. Soc.* **2008**, *130*, 408–409.

(41) Chiti, F.; Dobson, C. M. Amyloid formation by globular proteins under native conditions. *Nat. Chem. Biol.* **2009**, *5*, 15–22.

(42) Nelson, S. R.; Lyon, M.; Gallagher, J. T.; Johnson, E. A.; Pepys, M. B. Isolation and characterization of the integral glycosaminoglycan constituents of human amyloid A and monoclonal light-chain amyloid fibrils. *Biochem. J.* **1991**, *275*, 67–73.

(43) Blancas-Mejía, L. M.; Hammernik, J.; Marin-Argany, M.; Ramirez-Alvarado, M. Differential effects on light chain amyloid formation depend on mutations and type of glycosaminoglycans. *J. Biol. Chem.* **2015**, *290*, 4953–4965.

(44) Lu, J. X.; Qiang, W.; Yau, W. M.; Schwieters, C. D.; Meredith, S. C.; Tycko, R. Molecular structure of β-amyloid fibrils in Alzheimer's disease brain tissue. *Cell* **2013**, *154*, 1257–1268.

(45) Hrncic, R.; Wall, J.; Wolfenbarger, D. A.; Murphy, C. L.; Schell, M.; Weiss, D. T.; Solomon, A. Antibody-mediated resolution of light chain-associated amyloid deposits. *Am. J. Pathol.* **2000**, *157*, 1239–1246.

(46) Wasmer, C.; Lange, A.; Van Melckebeke, H.; Siemer, A. B.; Riek, R.; Meier, B. H. Amyloid fibrils of the HET-s(218–289) prion form a beta solenoid with a triangular hydrophobic core. *Science* **2008**, *319*, 1523–1526.

(47) Chiti, F.; Dobson, C. M. Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.* **2006**, *75*, 333–366.

(48) Solomon, A.; Weiss, D. T.; Wall, J. S. Immunotherapy in systemic primary (AL) amyloidosis using amyloid-reactive monoclonal antibodies. *Cancer Biother. Radio.* **2003**, *18*, 853–860.

(49) Marley, J.; Lu, M.; Bracken, C. A method for efficient isotopic labeling of recombinant proteins. *J. Biomol. NMR* **2001**, *20*, 71–75.

(50) Sambrook, J.; Russell, D. W.; Irwin, N.; Janssen, K. A. In *Molecular Cloning: A Laboratory Manual*; 3 ed.; Argentine, J., Ed.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York, 2001; Vol. 3, pp A2.2.

(51) Blancas-Mejía, L. M.; Tischer, A.; Thompson, J. R.; Tai, J.; Wang, L.; Auton, M.; Ramirez-Alvarado, M. Kinetic control in protein folding for light chain amyloidosis and the differential effects of somatic mutations. *J. Mol. Biol.* **2014**, *426*, 347–361.

(52) Pras, M.; Schubert, M.; Zuckerfr.D; Rimon, A.; Franklin, E. C. Characterization of soluble amyloid prepared in water. *J. Clin. Invest.* **1968**, *47*, 924–933.

(53) Van Geet, A. L. Calibration of the methanol and glycol nuclear magnetic resonance thermometers with a static thermistor probe. *Anal. Chem.* **1968**, *40*, 2227–2229.

(54) Comellas, G.; Lopez, J. J.; Nieuwkoop, A. J.; Lemkau, L. R.; Rienstra, C. M. Straightforward, effective calibration of SPINAL-64 decoupling results in the enhancement of sensitivity and resolution of biomolecular solid-state NMR. *J. Magn. Reson.* **2011**, *209*, 131–135.

(55) Takegoshi, K.; Nakamura, S.; Terao, T. 13C-1H dipolar-assisted rotational resonance in magic-angle spinning NMR. *Chem. Phys. Lett.* **2001**, *344*, 631–637.

(56) Baldus, M.; Petkova, A. T.; Herzfeld, J.; Griffin, R. G. Cross polarization in the tilted frame: assignment and spectral simplification in heteronuclear spin systems. *Mol. Phys.* **1998**, *95*, 1197–1207.

(57) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. NMRPipe: a multidimensional spectral processing system based on unix pipes. *J. Biomol. NMR* **1995**, *6*, 277–293.

(58) Goddard, T. D.; Kneller, D. G. SPARKY 3. University of California, San Francisco, **2008**.

(59) Hou, G.; Lu, X.; Vega, A. J.; Polenova, T. Accurate measurement of heteronuclear dipolar couplings by phase-alternating R-symmetry (PARS) sequences in magic angle spinning NMR spectroscopy. *J. Chem. Phys.* **2014**, *141*, 104202.

CHAPTER 4: Distinct Structural Features of Light-Chain Fibrils Revealed by Four-Dimensional ^{13}C-Detected Solid-State NMR Spectroscopy

4.1. NOTES AND ACKNOWLEDGEMENTS

Dennis W. Piehl, Alexander M. Barclay, Joseph M. Courtney, Marina Ramirez-Alvarado, Chad M. Rienstra. This research is supported by the University of Illinois (Centennial Scholars Award to C.M.R.). D.W.P. was an American Heart Association Predoctoral Fellow (15PRE25100008). We would like to thank the NMRBox team—particularly, Prof. Jeffrey C. Hoch, Dr. Mark W. Maciejewski, and Dr. Adam D. Schuyler—and service for their support in the management and processing of several non-uniformly sampled four-dimensional SSNMR data sets described in this chapter. We also extend gratitude to the AL patients and their families for their generous support of this research.

4.2. ABSTRACT

In this work, the structural features of light chain amyloidosis (AL) fibrils composed of the patient-derived protein AL-09 were investigated by the application of ^{13}C-detected four-dimensional heteronuclear correlation SSNMR experiments. In addition to enabling the efficient extension and confirmation of ^{13}C and ^{15}N backbone and side-chain chemical shift assignments for the fibril structure, our results reveal a distinct set of backbone torsion angle restraints for the central regions of the AL-09 variable domain (V_L) sequence, which were previously unattainable by traditional three-dimensional methods. SSNMR data with sufficiently high resolution and sensitivity for performing detailed spectral analyses were obtained through the implementation of band-selective J-decoupling pulses on the indirect ^{13}C-dimensions in $^{13}C\alpha_i$-$^{15}N_i$-$^{13}C'_{i-1}$-$^{13}CX_{i-1}$ 4D type experiments in combination with non-uniform sampling (NUS) regimes. Overall, our data suggest the existence of a unique structural motif for the core-forming residues of AL-09 V_L fibrils, shedding light on the interplay of the terminal and central regions of the sequence on the mechanisms of fibrillogenesis occurring in light chain amyloidosis pathogenesis.

4.3. INTRODUCTION

Light-chain (AL) amyloidosis is an aggressive systemic disease in which an over-expression of unstable immunoglobulin light-chain proteins in patients culminates in the formation and deposition of insoluble fibrils at various sites throughout the body, including the heart, kidneys, liver, and peripheral nerves[1,2]. The median survival for AL patients is 12–40 months, and becomes even more severe when the afflicted organ is the heart, which occurs in more than 48% of patients[1-4]. There is currently no known cure for AL, and efforts towards identifying effective diagnostic and treatment strategies, such as structure-based drug discovery, would largely benefit from a greater understanding of the molecular determinants that drive the fibril formation process as well as the structural details of the toxic fibril species[5-8].

Molecular-level insight into the structure of insoluble biomolecules such as protein fibrils can be obtained through the use of solid-state nuclear magnetic resonance (SSNMR) spectroscopy. Previous studies by our group and others have demonstrated the application of two- and three-dimensional SSNMR experiments to the study of AL fibrils, providing the first sets of high resolution details on the secondary structure and dynamics for two separate AL patient-derived light-chain sequences[9,10]. However, due to the observation of partial ^{13}C and ^{15}N chemical shift degeneracy (e.g., arising from repetitive amino acid motifs in the sequence) or weaker signal intensity for slightly more dynamic regions of the fibril structure, the ability to uniquely resolve and sequentially assign all of the structured regions of the fibrils—and thus arrive at a complete three-dimensional structure—has remained an inherent challenge.

Indeed, a vast number of effective methods have been developed for addressing such limitations on spectral resolution in SSNMR studies of biomolecules. One of the most direct approaches that researchers have pursued is the use of higher-dimensional internuclear correlation experiments (e.g., 4D and 5D spectroscopy), which can allow the removal of chemical shift degeneracy that may be present in 3D spectra. For example, SSNMR experiments have been developed to employ inter-residue polarization transfer schemes, such as the CANCOCX 4D sequence[11], which can been applied to insoluble protein systems to yield sequential, site-specific chemical shift assignments of backbone atom correlations[12].

Additionally, the removal of homonuclear ^{13}C-^{13}C J-couplings can serve as an alternative means to improving spectral resolution, either through the application of band-selective pulses on uniformly [^{13}C, ^{15}N]-labeled samples[13,14] or the preparation of sparsely ^{13}C-labeled protein samples using [1,3-^{13}C]- or [2-^{13}C]-labeled glycerol as a growth substrate[15]. In effect, both strategies offer substantial enhancements to the spectral resolution of unique internuclear cross peaks, thus relieving limitations to spectral analysis due to chemical shift degeneracy.

Importantly, however, achieving the full benefit on resolution through these methods also typically requires substantial lengthening of the total experiment time, sometimes even on the order of multiple weeks. As such, the combination of these techniques with more efficient data acquisition schemes—particularly non-uniform sampling (NUS)—can serve as a powerful experimental strategy for achieving the high-resolution necessary for resolving otherwise overlapped cross peaks in a practical amount of time[16]. The use of NUS in 4D SSNMR experiments is becoming increasingly more routine among researchers as a strategy for effectively circumventing this time-constraint, in which a much smaller percentage of the multidimensional sampling grid (in some cases as little as 1%) may be acquired and still provide the sensitivity and resolution necessary for complete spectral analysis[17]. For instance, previous SSNMR investigations have demonstrated the utility of NUS 4D spectroscopy to obtain ^{1}H-^{1}H distance restraints through ^{1}H-^{15}N-^{15}N-^{1}H-edited 4D experiments[18]. Furthermore, this approach has been successfully applied for improving the extent of chemical shift assignments for insoluble protein systems by performing ^{1}H-detected NUS 4D experiments to obtain intra- and inter-residue backbone correlations[19] as well as ^{13}C-detected dynamic nuclear polarization (DNP) NUS 4D experiments to achieve protein side-chain correlations[20].

Here we describe the implementation of 4D NUS ^{13}C-detected heteronuclear correlation experiments in combination with selective pulses to achieve ^{13}C-^{13}C scalar decoupling and SPC-N mixing to enable the efficient acquisition of spectral data with substantial improvements in resolution and sensitivity. We apply this to the case of AL-09 V_L fibrils to allow for the nearly complete chemical shift assignments of the observed structured fibril regions. Our results reveal that the central region of the

protein sequence exhibits distinct conformational and dynamical characteristics relative to other regions of the sequence, and may indicate the existence of a loop structure within the rigid fibril core that is primarily stabilized by the highly immobile residues of the N- and C-termini. These insights strongly support the hypothesis that the non-conservative mutations are responsible primarily for destabilizing the dimer complex and allowing misfolding to occur rather than being involved structurally in the final fibril form, while the N- and C-terminal residues (with partial involvement from the other central regions of the sequence) are integral for maintaining the structural rigidity and integrity of the AL-09 V_L fibril species.

4.4. RESULTS AND DISCUSSION

4.4.1. Achieving resolution and sensitivity enhancements through J-decoupled ^{13}C-detected 4D SSNMR spectroscopy.

We present the implementation of a 4D pulse sequence that combines options for J-decoupling pulses and SPC-N mixing with ^{13}C-detection, allowing for highly efficient polarization transfer for both $^{13}C\alpha_i$-$^{15}N_i$-$^{13}C'_{i-1}$-$^{13}CX_{i-1}$ (CANCOCX; *i* to *i*-1) or $^{13}C'_{i-1}$-$^{15}N_i$-$^{13}C\alpha_i$-$^{13}CX_i$ (CONCACX; *i*-1 to *i*) correlation pathways (Figure 4.1A). The incorporation of these elements is illustrated in the pulse sequence diagrams in Figures 1B-C. The sequence begins with the transfer of initial magnetization from ^{1}H to ^{13}C by SPECIFIC cross-polarization (CP), followed by the first chemical shift evolution period (t_1)[21]. Polarization is then transferred to ^{15}N for the second evolution period (t_2) and back to ^{13}C for the third indirect dimension (t_3), followed by a ^{13}C-^{13}C mixing period (either DARR or SPC-N)[22,23] and finally direct detection of the resulting signal (t_{acq}). SPECIFIC-CP is used for all heteronuclear transfer steps[21], during which continuous-wave (CW) decoupling is applied on the ^{1}H channel (apart from the ^{1}H-^{13}C transfer step); SPINAL-64 is used for ^{1}H decoupling during all evolution periods and acquisition[24]. In addition, band-selective (soft) pulses may be applied to both indirect ^{13}C dimensions[14,25] to provide further resolution and sensitivity enhancements (during which TPPM is used for ^{1}H-decoupling)[26]. Last, to enable the efficient acquisition of 4D data sets, we digitized all three indirect dimensions according to a NUS regime with sampling densities of 0.5–3.7% and exponential decay rates of 4–7 Hz[16,27].

We evaluated the improvement in resolution and sensitivity upon additional implementation of J-decoupling soft pulses by performing our initial experiments on the β1 immunoglobulin binding domain of protein G (GB1) as a protein standard. Towards this end, following the acquisition of both CANCOCA SPC-5 q60 (1.08 ms) and CANCOCX 50 ms 4D spectra with and without the use of soft pulses (Table 4.1), respectively, we analyzed and compared the line widths of the indirect ^{13}C dimensions and signal-to-noise ratios (SNR) of all 55 $^{13}C\alpha_i$-$^{15}N_i$-$^{13}C'_{i-1}$-$^{13}C\alpha_{i-1}$ peaks observed in the data sets[28]. Overall, this revealed an average improvement in line width of 45 Hz and 43 Hz ± 10 Hz for the ω1 and ω3 indirect ^{13}C-dimensions, respectively. The removal of J-couplings via application of soft pulses represents the primary contributor to improvements in resolution, as is supported by the analysis of changes in line widths observed for the ^{15}N dimension (on which no selective pulse is applied) resulting from the use of a 50% longer evolution time, which revealed a 7 Hz ± 10 Hz reduction line width. We also note that the use of slightly longer evolution times in combination with the use of soft pulses not only provides an additional enhancement to the resolution in the indirect ^{13}C dimensions, but also allows for the full benefit of sensitivity to be achieved by removing the $\cos(\pi Jt)$ modulation from the signal decay function for the indirect dimension evolution period[29]. Accordingly, the average sensitivity of cross peaks observed in the experiment with applied soft pulses remained within 8% of the average observed in the experiment without soft pulses. Thus, these results demonstrate the capacity of the presented CNCC 4D experiment to provide substantial improvements to spectral resolution while at the same time retaining maximal sensitivity in cross-peak intensity.

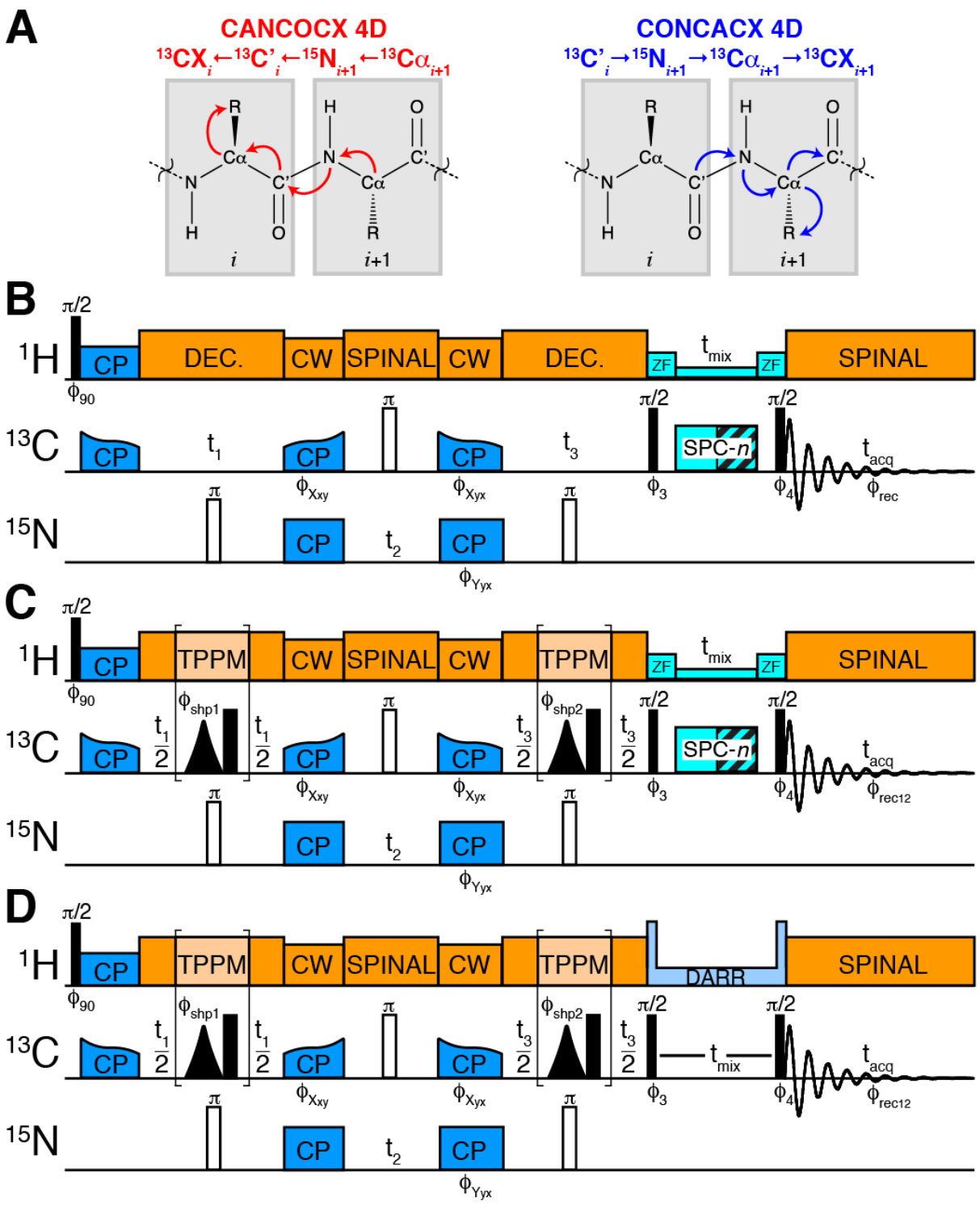

Figure 4.1. (A) Polarization transfer pathway schematic and (B–C) variety of ^{13}C-^{15}N-^{13}C-^{13}C 4D pulse sequence setups performed on U-^{13}C,^{15}N labeled AL-09 V_L fibrils to obtain a diverse set of spectral information for enabling chemical shift and structural analyses.

4.4.2. Chemical shift assignments by CNCC 4D experiments reveal that central regions of the AL-09 V_L sequence are involved in the fibril core.

We proceeded to perform several variations of these ^{13}C-^{15}N-^{13}C-^{13}C 4D experiments on U-[^{13}C,^{15}N] labeled AL-09 V_L fibrils, in order to provide inter-residue backbone and sidechain correlation information that complements the 3D data reported previously[30]. In particular, as detailed in Table 4.1, these included: a CANCOCA SPC-8 q64 (1.28 ms) 4D without J-decoupling and with modest digitization lengths ($t_1$max = 5.12 ms, $t_2$max = 10.24 ms, $t_3$max = 5.12 ms) (Figure 4.1B); a CANCOCA SPC-8 q64 4D with J-decoupling pulses and extended digitization times ($t_1$max = 7.68 ms, $t_2$max = 15.36 ms, $t_3$max = 10.24 ms) (Figure 4.1C); and a CANCOCX 50 ms (DARR) 4D under the same conditions (Figure 4.1D), allowing for the observation of *i* to *i*-1 inter-residue correlations with polarization transfer optimized for the emphasis of Cα or side-chain resonances in the ω4 dimension. Subsequently, the complementary CONCACO SPC-8 q64 4D spectrum was acquired with similar digitization for each corresponding nucleus and the application of J-decoupling soft pulses to achieve *i* to *i*+1 inter-residue backbone connectivity. In order to enable the efficient collection of these data sets with optimal digital resolution, all 4D experiments were performed according to a 2.0–3.7% NUS schedule with a 4–7 Hz exponential decay biasing factor and digitizing out to maximum evolution times near the T2 value of the corresponding nucleus. For instance, 4D experiments employing soft pulses on both indirect ^{13}C-dimensions were collected using maximum evolution times of 7.68 ms, 15.36 ms, and 10.24 ms for $^{13}C\alpha$, ^{15}N, and $^{13}C'$, respectively. As the use of these digitization limits in combination with sufficiently large spectral widths would demand in the uniformly-sampled case approximately 10 weeks of continuous data collection (~68.5 days for 73,728 total rows), the execution of these experiments with only 2.0% coverage allowed for the experiment time to be reduced to less than 1.4 days (1472 rows). (For the CANCOCA SPC-8 q64 4D experiment without applied soft pulses, which used shorter digitization times and a 0.5 s longer pulse delay, the use of a 3.7% NUS schedule reduced the experiment time from 49 days to 1 day.)

Analysis of the resulting CANCOCA/CX and CONCACO 4D spectra revealed strong and well-resolved cross peaks with chemical shift frequencies in direct agreement with resonance assignments made previously with 2D and 3D spectra[30]. Notably, these 4D spectra revealed the presence of approximately 85 spin systems with high cross-peak intensity (i.e., greater than 7 times the noise floor), based on the number of $^{13}C\alpha_i$-$^{15}N_i$-$^{13}C'_{i-1}$-$^{13}C\alpha_{i-1}$ peaks reconstructed by SMILE[31] for the CANCOCA 3.7% NUS 4D data set. Accordingly, this signifies that the number of highly ordered residues in the AL-09 V_L fibril structure is closer to the upper limit of our estimation based on 2D and 3D spectra alone, which was previously reported to be 70 ± 15 residues[30]. This apparent increase in the number of structured residues for AL-09 V_L fibrils may be attributed to the further removal of spin-system degeneracy and peak overlap that were previously observed in 3D spectra through the addition of a fourth-dimension.

Subsequent analysis of the observed sets of cross peaks in the 4D spectra allowed for the identification and unequivocal confirmation of all 48 spin-system assignments reported in prior studies[9,30], many of which had previously presented notable assignment challenges due to extensive chemical shift degeneracy. An example of the removal of this degeneracy by extending the dimensionality of our data sets to 4D spectra is illustrated in Figure 4.2. As portrayed along the dashed-line through the CAN(co)CX 3D strip in Figure 4.2A, two spin-systems—V15Cα-N-S14C'/CX and I21Cα-N-T20C'/CX—share nearly identical $^{13}C\alpha$, ^{15}N, and $^{13}C'$ chemical frequencies and were previously only able to be separated through the subsequent acquisition and analysis of other 3D spectra[30]. However, as shown in the 2D plane of the CANCOCX 50 ms 4D in Figure 4.2B (corresponding to $^{13}C\alpha$ (F1) of 60.8 ppm and ^{15}N (F2) of 129.5 ppm), the introduction of an additional $^{13}C'$ (F3) dimension reveals a clear separation of the two spin-systems by approximately 0.4 ppm.

The reduction of chemical shift degeneracy observed in the 4D spectra also enabled efforts towards extending the backbone and side-chain resonance assignments for AL-09 V_L fibrils. A clear demonstration of this ability can be observed for the case of P80–E81, in which the large extent of

degeneracy in the carbonyl region of the CAN(co)CX 3D (which prevented the identification of the P80C' and Cα resonances previously; Figure 4.3A) is removed in the CANCOCA SPC-8 q64 4D spectrum, revealing a single, uniquely resolved cross peak corresponding to both the C' and Cα chemical shifts of P80 (Figure 4.3B). Furthermore, upon the application of J-decoupling soft pulses in the two indirect ^{13}C-dimensions (F1 and F3), an additional improvement to resolution is observed in the reduction of line width by approximately 50 Hz (Figure 4.3C).

Collectively, our analysis of the acquired CNCC 4D spectra—combined with the augmented interpretation of NCACX, NCOCX, and CANcoCX 3D data sets[9,30]—allowed for the assignment of resonances for 81 residues in the AL-09 V_L sequence (72 of which were identified in the 4D data alone). Of these 81 residues, the complete set of backbone chemical shifts were assigned for 67 residues, while partial backbone and/or side-chain assignments were made for an additional 14 residues (Figure 4.4). Importantly, along with providing a substantial improvement to the 43% of the sequence assigned previously[30], the expanded set of of backbone assignments presented here represents more than 90% of the complete set of chemical shifts for the observable residues comprising the structured core of AL-09 V_L fibrils (i.e., 81 assigned residues of the approximately 85 spin systems observed in the 4D data sets). Furthermore, two of the additional amino acid stretches that were assigned here include those from T56–E70 and S76–T85, both of which occur near the central region of the AL-09 V_L sequence. Thus, clearly the use of highly digitized 4D spectra has helped us assign unique regions of the fibril core, indicating that the core may also involve structural contributions from some of the central regions of the protein sequence, along with the N- and C-terminal regions that we originally identified based on 3D spectra[9].

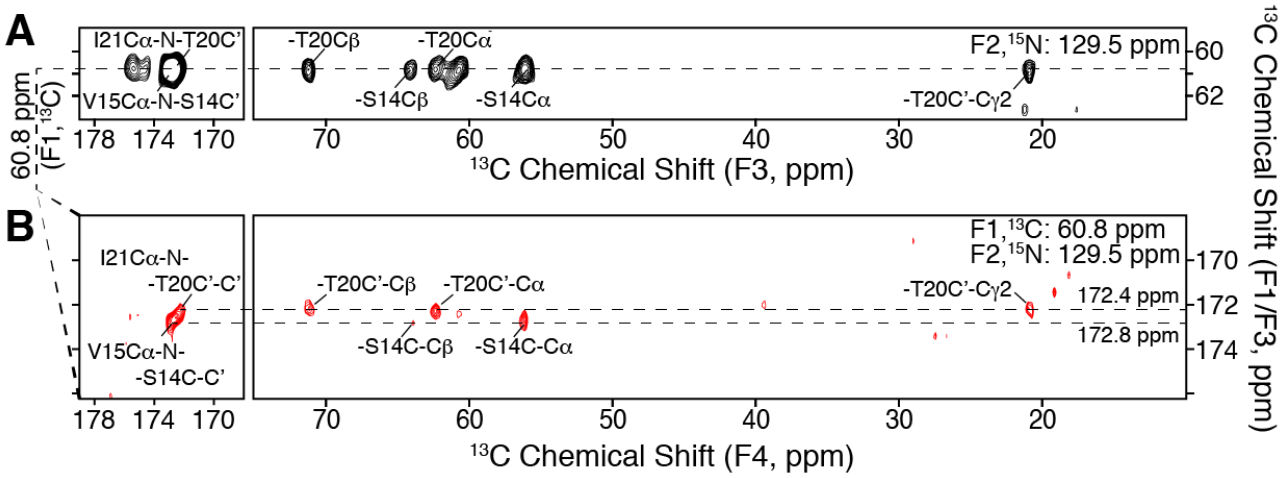

Figure 4.2. Example of removing chemical shift degeneracy of the S14–V15 and T20–I21 spin-systems between (A) $^{13}C\alpha_i$-$^{15}N_i$-($^{13}C'$)$_{i-1}$-$^{13}CX_{i-1}$ 3D and (B) $^{13}C\alpha_i$-$^{15}N_i$-$^{13}C'_{i-1}$-$^{13}CX_{i-1}$ 4D spectra, in which the S14- and T20-$^{13}C'$ frequencies differ by only 0.4 ppm.

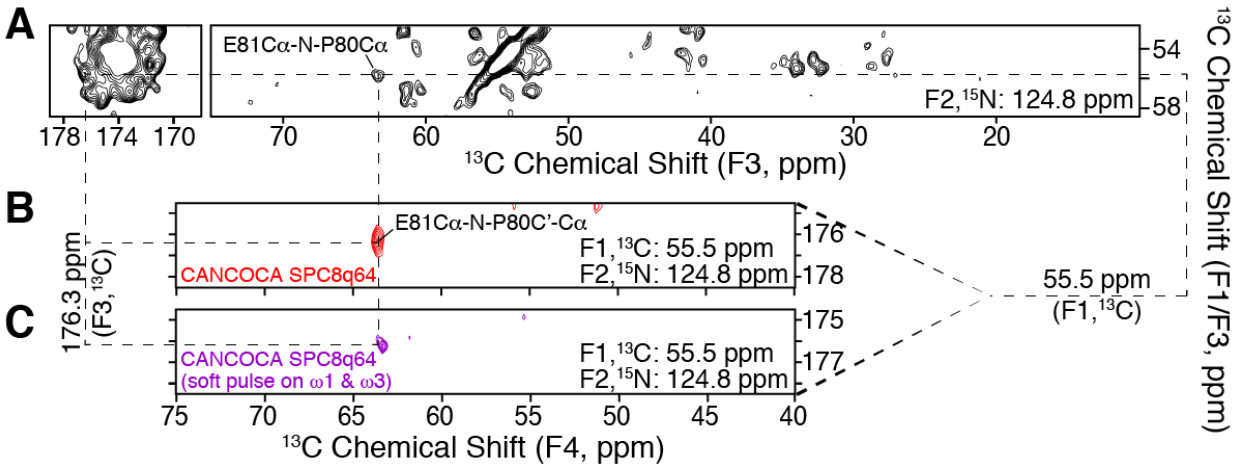

Figure 4.3. Identification of new chemical shift assignments for residues P80 and E81 in AL-09 V_L fibrils by extending (A) CAN(co)CX 25 ms 3D SSNMR experiments to (B) CANCOCA SPC-8 q64 (1.28 ms) 4D NUS experiments. (C) The application of J-decoupling soft pulses on both indirect ^{13}C-dimensions in (B) provides further resolution enhancements for internuclear correlation cross peaks.

4.4.3. Centrally-occurring residues exhibit a distinct conformation from the N- and C-terminal segments in AL-09 V_L fibrils.

The assignment of several previously unresolvable regions of the AL-09 V_L protein sequence allowed for the subsequent identification of several key insights into the secondary structure of the fibril species as well as indications of tertiary structural components. We first investigated these details through an analysis of the secondary ^{13}C chemical shift trends ($\Delta\delta\ ^{13}C$) for all assigned residues. As shown in Figure 4.4, the majority of the sequence exhibits average $\Delta\delta\ ^{13}C$ values that are consistent with a β-strand secondary structure, as is particularly evident for residues near the N- and C-termini (with the exception of a few deviations observed for residues S9, D28, and Q100–T102, likely signifying the occurrence of a β-hairpin or turn at that position of the structure). Additionally, the stretch of residues from S76–T85 demonstrates secondary shift trends indicative of primarily β-strand content. Interestingly, however, a notable divergence occurs for the residues spanning the central region of the sequence, T56–E70, in which no single secondary structure prediction is particularly dominant (Figure 4.4)[32,33]. Moreover, as may be evident by the large variability in secondary chemical shift across this region, several of the residues within this stretch are also observed to exhibit a relatively unique set of isotropic chemical shifts that may suggest the formation of a structural arrangement beyond simply random coil. For example, one particularly striking set of backbone chemical shifts observed are those of glycine residues in the sequence motif, $_{63}$SGSGSGT$_{69}$ (Figure 4.5), in which G64 and G68 demonstrate downfield ^{15}N and $^{13}C\alpha$ chemical shifts that are 1.5–1.7 standard deviations away from the statistical mean[34] (G64, 121.1 ppm and 49.2 ppm; G68, 122.0 ppm and 48.7 ppm), whereas the intervening residue, G66, exhibits a set of chemical shifts that are very close to the mean (G66, 109.6 ppm and 44.2 ppm). Accordingly, although no specific secondary structure may be immediately evident from these absolute chemical shift values, the relative site-to-site variation of this short motif may suggest the presence of a less common structural element, which we pursue further below.

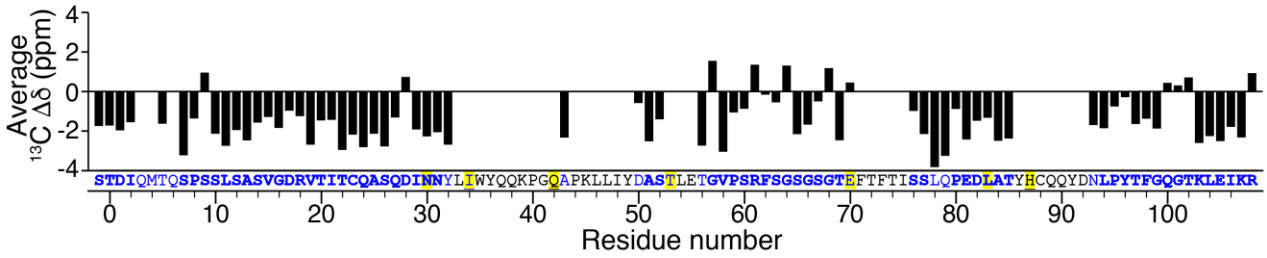

Figure 4.4. Extent of chemical shift assignments and corresponding average secondary ^{13}C chemical shift trends ($\Delta\delta$ ^{13}C)[32] as a function of residue number for AL-09 V_L fibrils. Residues labeled in blue are assigned; those in bold are unambiguous. Somatic mutations are highlighted; those which are underlined are non-conservative.

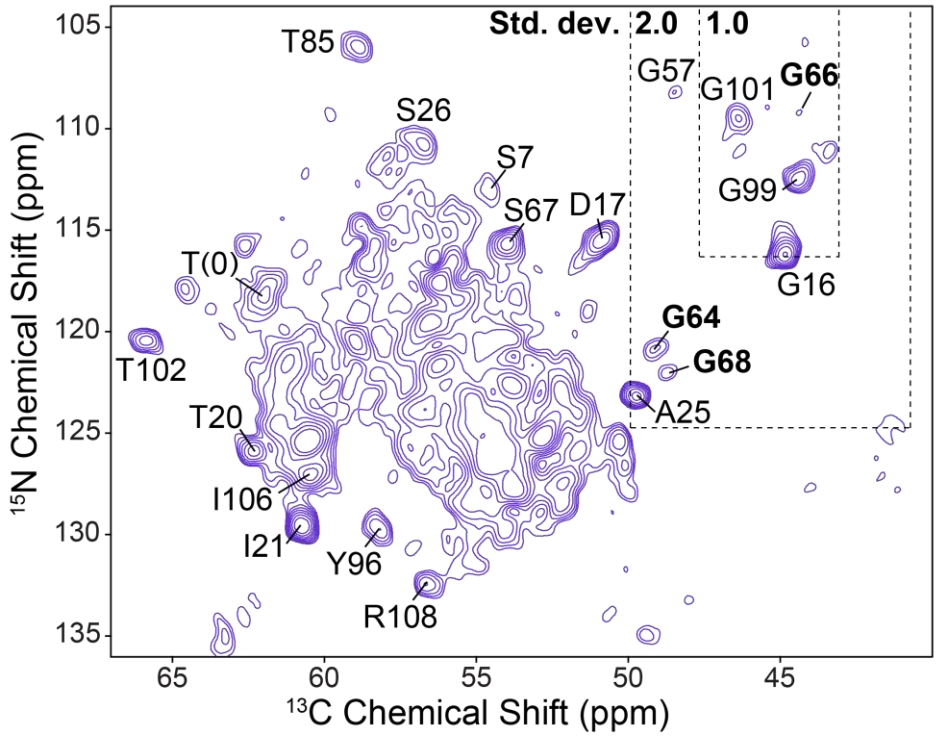

Figure 4.5. ^{15}N-^{13}Cα 2D spectrum of AL-09 V_L fibrils. Cross peaks corresponding to glycine residues, G64, G66, and G68 are emphasized with bold labels. Regions enclosed by the dashed lines represent the boundaries of 1.0 and 2.0 standard deviations away from the statistical mean of glycine ^{15}N and ^{13}Cα chemical shifts (data obtained from the BMRB[34]). Spectrum was acquired at 600 MHz (^{1}H frequency) on an InfinityPlus spectrometer with a MAS rate of 13.333 kHz and variable temperature set point of -10 °C.

To further investigate the structural implications of the secondary chemical shift trends on the AL-09 V_L fibril form, we proceeded to analyze chemical shift-based dihedral angle predictions using TALOS-N[35], providing an extensive list of backbone and sidechain torsion angle restraints (Figure 4.6). The results from these predictions reveal that residues from T56–E70 exhibit a generally mixed set of torsion angles—consistent with our observation of the varying distribution of secondary chemical shifts in this region—while residues S76–T85 demonstrate torsion angles that are generally consistent with a β-strand secondary structure. TALOS-N was unable to make predictions for several of the residues in the T56–E70 stretch, primarily for the Gly residues; yet, we assume this is due to a combination of the limited number of training sets available for Gly motifs as well as the uniqueness of chemical shift values and amino acid pattern within the sequence. Several prior studies have identified similar shortcomings in the TALOS-N database for Gly residues in fibrils[36-39]. As such, the secondary structure of these residues cannot be reliably predicted from TALOS-N.

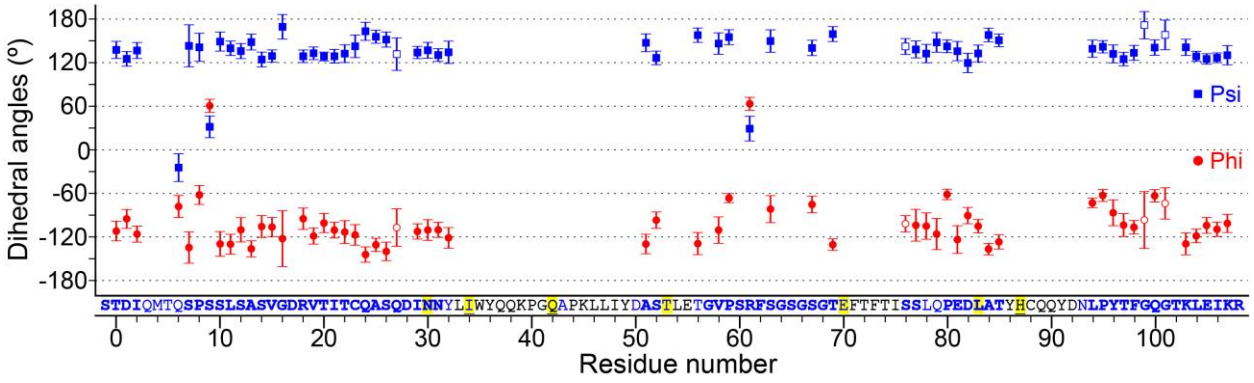

Figure 4.6. TALOS-N[35] predicted backbone torsion angles for AL-09 V_L fibrils as a function of residue number. Blue squares and red circles indicate psi and phi torsion angles, respectively. Filled and open points were reported as "strong" or "generous" predictions, respectively. Angles that were reported as "warn" predictions are not shown. Error bars represent the standard deviations reported by TALOS-N.

In light of this predictive limitation, we inquired on how the backbone dihedral angles of the X-ray dimer structure of AL-09 V_L (PDB: 2Q1E)[40] compare to those which were able to be predicted by TALOS-N, in order to possibly shed light on potential similarities or differences that may inform on the gaps in predictions for the fibril structure of AL-09 V_L (Figure 4.7). As the soluble dimer form of AL-09 V_L is known to contain a high degree of β-sheet content (as is the case for all native light chain protein structures), many of the TALOS-N predicted angles for the fibril structure are qualitatively similar for regions of the sequence near the N- and C-termini. This observation may suggest that a subset of the β-strands that are present in the dimer form may also exist with similar chain lengths in the fibril state, although likely in a different orientation and with different intramolecular interactions given the notable differences observed for dihedral angles at different regions of the sequence. A relatively larger number of differences in torsion angles are observed for the central sequence region of T56–T69 (apart from the gaps in predictions discussed above), in which only two of the seven residues that were able to be compared demonstrated qualitative agreement in both phi and psi angles between the crystal structure and fibril predictions, V58 and P59. While we cannot completely eliminate the possibility of a similar structural motif occurring in the fibril form of AL-09 V_L at this region of the sequence given the gaps in predicted angles, the general lack of agreement for the majority of comparable residues in the central region makes this possibility seem unlikely.

One interesting detail about the X-ray dimer structure worth noting is that within the S63–T69 stretch, residues G66 and G68 both exhibit positive phi angles of around 90°; yet, both residues exist in a β-strand in the 3D structure, which immediately leads into a turn and an adjacent β-strand[40]. This insight further emphasizes the difficulty associated with accurately predicting the conformation of glycine residues due to the highly flexible nature of the backbone, and becomes even more challenging in the case of a fibril structure which comprise less than 0.05% of the number of protein structures available on the RCSB PDB[41]. As such, we decided to pursue a comparison of the AL-09 V_L fibril chemical shifts with chemical shifts observed for other fibril structures with similar amino acid sequences. Although the

majority of fibril structures that have been investigated by SSNMR methods contain a high number of hydrophobic residues—which differ from the largely polar-containing residues in AL-09 V_L fibrils—several recent studies offer particularly suitable candidates for comparison.

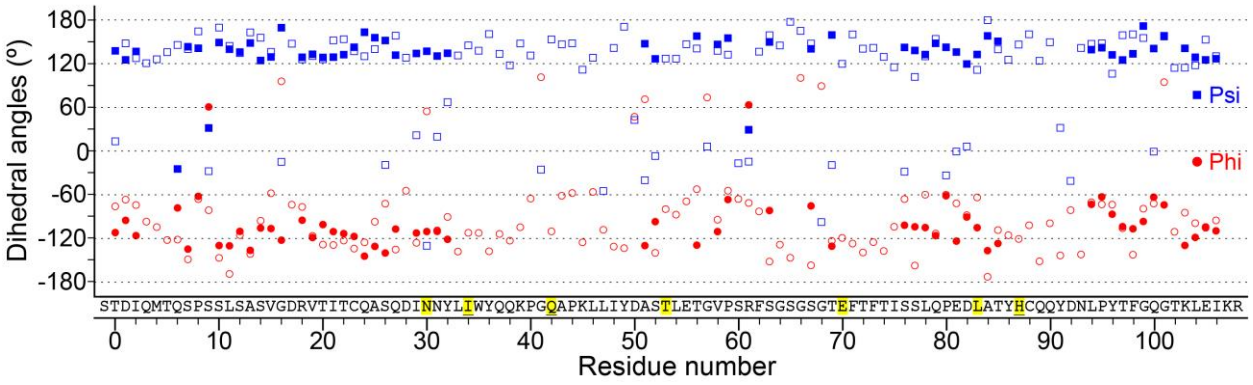

Figure 4.7. Comparison of backbone torsion angles corresponding to the X-ray crystal structure of the AL-09 V_L dimer (PDB: 2Q1E; open symbols)[40] and those determined from SSNMR-based TALOS-N predictions of AL-09 V_L fibrils (closed symbols). Psi angles are depicted as blue squares (open or closed); phi angles are depicted as red circles (open or closed).

The first is the case of MAK33 S20N V_L fibrils, which represent another instance of AL disease-related amyloid fibrils[10]. As the MAK33 V_L domain is derived from the same kappa germline gene family as AL-09 V_L, and thus shares a high sequence identity of 57% overall, including the same stretch of residues between P59–T69[10]. Interestingly, however, the chemical shifts observed for residues G64, G66, and G68 in AL-09 V_L fibrils are largely disparate compared to those for MAK33 S20N V_L fibrils, in which all assigned Gly residues exhibit ^{15}N and $^{13}C\alpha$ frequencies falling below 112 ppm and 47 ppm, respectively. Despite both AL-09 and MAK33 V_L domains being derived from the kappa family of V_L germline genes and sharing a large percentage of sequence identity, each protein appears to assume a rather different fibril core structure. This notable difference was originally proposed by Hora et al.

through their observation and assignments of primarily the central regions of the protein sequence as opposed to the terminal residues as observed for the case of AL-09 V_L[10,30].

We also compared our chemical shift and secondary structure observations with those for the recently reported structure of FUS-LC fibrils, which exhibit a similarly high composition of Gly, Ser, and Thr residues[38]. Although the FUS-LC and AL-09 V_L proteins essentially have no direct sequence identity, they both exhibit a large extent of interleaving polar and non-polar residues throughout the observable region of the fibril core[38]. Interestingly, the authors report the presence of a loop structure for a stretch of residues of the sequence, $_{74}$GYGSTGGYGSSQSS$_{87}$, in which the polar Ser, Thr, and Tyr residues form a hydrogen-bonding network proposed to be responsible for holding the loop together[38]. In comparison to the T56–T69 stretch in AL-09 V_L fibrils (with the sequence, $_{56}$TGVPSRFSGSGSGT$_{69}$), a disproportionately high number of Gly, Ser, and Thr residues are observed in both cases. Furthermore, the authors were unable to report TALOS-N dihedral angle predictions for 8 of the residues in this stretch of residues for FUS-LC, including G82 which exists at the sharp turn at the edge of the loop and exhibits the most downfield ^{13}Cα chemical shift of all observable Gly residues in the FUS-LC core of ~49 ppm[38]. In light of the similarities in chemical shift and neighboring residue identity, it thus remains highly possible that a similar conformation and environment is true for the case of the G64 and G68 residues in AL-09 V_L fibrils, in which each residue forms part of a greater loop motif in the core.

Another very striking set of chemical shifts is presented by the nearby residue R61, in which the values for ^{13}Cβ (26.2 ppm) and ^{13}Cγ (32.6 ppm) are 1.7 standard deviations upfield and downfield from the statistical means, respectively[34]. Despite the occurrence of each resonance closer to the other's average value (i.e., the average shifts for Arg ^{13}Cβ and ^{13}Cγ are 30.7 ppm and 27.3 ppm, respectively), we were able to distinguish and unambiguously assign the appropriate chemical shift to the Cβ and Cγ atoms of R61 through the implementation of SPC-N mixing, as this provides a sign inversion for the signal arising from directly coupled nuclei relative to those that are indirectly coupled (i.e., a negative peak is observed for Cα-Cβ correlations whereas a positive peak is seen for Cα-Cγ correlations[23]. While the

observation of Arg $^{13}C\beta$ chemical shifts around 26 ppm is not uncommon (~350 counts reported to the BMRB), the frequency of $^{13}C\gamma$ shifts between 32–33 ppm is considerably lower (< 50 counts per 0.1 ppm interval)[34]. Together, these upfield and downfield chemical shift deviations for the $^{13}C\beta$ and $^{13}C\gamma$ atoms of the Arg side chain, respectively, may indicate the adoption of a helical conformation which is distinct from the primarily β-sheet secondary structure observed for the other regions of the sequence, in addition to the formation of highly polar or electrostatic interactions with the side chains of other nearby residues.

One example of a similarly downfield-shifted Arg $^{13}C\gamma$ resonance was observed for the disulfide-stabilized dimer structure of the XCL1 chemokine in solution, exhibiting a $^{13}C\gamma$ shift of 32.9 ppm for residue Arg61 (referred to here as Arg61 to distinguish from R61 in AL-09 V_L) (BMRB: 25693)[42]. The structure of this dimer (PDB: 2N54)[42] reveals that Arg61 exists on an unstructured coil running alongside a β-strand of predominantly polar residues, including three Thr and two Tyr residues situated within 5 Å across from Arg61. These features have several implications on the possible influence on the side-chain chemical shifts of the Arg residue. First, the presence of two aromatic rings from proximal Tyr residues may subject Arg61 to ring-current effects, which are well-known to induce chemical shift perturbations on the order of several ppm for nearby ^{13}C resonances[43,44]. Second, the high number of nearby polar residues in the structure may induce substantial changes in the electrostatic environment surrounding the Cβ and Cγ side-chain atoms of Arg61, an effect well known to lead to large chemical shift changes[45].

In a strikingly similar fashion, R61 in AL-09 V_L is positioned in a stretch of amino acids given by, $_{56}$TGVPS<u>R</u>FSGSGSGT$_{69}$, in which it is immediately neighboring a polar Ser and nonpolar aromatic Phe residue, followed by an extensive sequence of generally intervening polar and non-polar amino acids (including two additional Phe residues at positions 71 and 73). Furthermore, the amino acid composition around R61 is rather unique relative to the other two Arg residues present in the sequence, R18 and R108, in which the closest aromatic residues to each are 14 and 12 aa away, respectively. Neither R18 nor R108 demonstrate similar chemical shift deviations as observed for R61. Accordingly, both the presence of polar and aromatic residues near R61 may be responsible for inducing the dramatic Cβ and Cγ chemical

shift perturbations that are observed, via a combination of ring current effects as well as contributions from electrostatics. Additionally, given that the Cβ exhibits an upfield shift relative to the average while the Cγ exhibits a downfield shift, it may be possible that the ring current from F62 produces an electric field gradient that affects both carbon atoms differently. Interestingly, TALOS-N χ^1 predictions[35] for the sidechain torsion angles of AL-09 V_L fibrils suggest that the R61 N-Cα- C β- Cγ dihedral angle assumes a gauche(-) rotamer conformation, with a χ^1 angle of -63° ± 12°. This may represent the geometry that is necessary for supporting the complex interactions and electrostatic environment that are experienced by the R61 side chain. One final interesting similarity between the XCL1 protein and AL-09 V_L, is that one of the main activities of the dimer form for XCL1 (as opposed to the monomer form) is the binding of GAG molecules[42].

Together, this set of unusual chemical shifts observed for the S60-T69 residue stretch of the AL-09 V_L sequence suggests that this region assumes a unique conformation within the fibril core relative to the N- and C- terminal regions. Notably, as observed by the ^{15}N and $^{13}C\alpha$ chemical shifts of the three Gly residues in this stretch ($_{63}$SGSGSGT$_{69}$), those observed for G64 and G68 are highly distinct from those of G66 (^{15}N, 109.7 ppm; $^{13}C\alpha$, 44.2 ppm) despite the generally repetitive amino acid sequence. A chemical shift variation of this magnitude between the central residue G66 and the flanking residues G64 and G68 may be explained the adoption of a β-turn or β-loop structure at this region of the AL-09 V_L sequence. In particular, the existence of a double β-turn can be rationalized through the results from previous solid-state studies on the structure of *Bombyx mori* silk fibroin, formed from the repetitive sequence motif of $(AGAGAG)_{10}$, which is presented to adopt an extended chain of repeating type-II β-turns maintained by hydrogen bonds between the carbonyl group of Gly_i with the amide hydrogen of Ala_{i+3}[46,47]. Additionally, the alternating direction of each consecutive type-II β-turn results in the placement of every other Gly residue (i.e., Gly_i and Gly_{i+5}) on the same face of the fiber chain and with an identical geometry[46]. Such a model would be consistent with the presented data for AL-09 V_L fibrils, as the high similarity in ^{15}N and ^{13}C chemical shifts for G64 and G68 (as opposed to G66) implies that the two residues exist in similar

chemical environments and orientations, and are possibly exposed to the same part of the fibril core or to the bulk-solvent. Further, this arrangement is also shown to support the formation of hydrogen bonds between stacked chains through C-O and N-H interactions[48], a feature which is fundamental to the cross-β structure formed by amyloid fibrils at large[49] and thus would be critical for maintaining intermolecular packing between AL-09 V_L monomers along the fibril axis. In addition to the backbone interaction, the presence of both interleaving Ser and Thr residues along this region may contribute further to the hydrogen bonding network responsible for maintaining the structural integrity of this region of the fibril core.

An alternative model would be the formation of a β-loop type structure, beginning with S60 and ending with T69. This would place residues G64 and G68 near the base of the loop (where it begins and ends), and G66 would be centered at the opposite end of the loop. As the nature of the chemical environment and types of interactions existing between these two ends of the loop would differ, such a motif may account for the different set of chemical shifts observed for G64 and G68 relative to G66. Moreover, given the alternating pattern of polar and nonpolar residues over this stretch, the Ser and Thr sidechains will face toward the same direction in the three-dimensional structure. Thus, the formation of a loop would be supported by the positioning of the hydroxyl side chains on the inside of the loop, in which the polar interactions between each residue provide an additional form of structural integrity to the fibril core. Further, the residues just outside of the loop—namely, S60-R61-F62—may represent another sharp turn in the arrangement of the core, consequently lending contribution to the aberrant chemical shifts observed for these residues of the AL-09 V_L fibril structure.

4.4.4. AL-09 V_L fibrils demonstrate subtle variations in dynamics throughout the extended core.

In addition to providing structural insight on the central regions of the AL-09 V_L sequence in the fibril core, the new assignments enabled by the four-dimensional J-decoupled CNCC experiments allowed for the subsequent evaluation of the corresponding dynamics relative to other segments of the

core. In particular, because the presented 4D experiments were performed using a dipolar-based polarization transfer scheme (i.e., cross-polarization) to achieve heteronuclear correlations, we first assessed the relative peak heights of each CANCOCA peak to gain a qualitative insight on the relative dynamics of different regions of the protein sequence (Figure 4.8). This analysis reveals a substantially reduced signal intensities measured for residues T56-E70 compared to those for the stretch from S7-I29 (N-terminal residues), and a slight reduction in peak height relative to residues L94-R108 (C-terminal residues), indicating that the residues in the central region of the sequence are generally more dynamic (Figure 4.8). Additionally, because the 4D experiments allowed for the separation of peaks that are overlapped in CAN(co)CX 3D experiments, this analysis also provides a more complete and representative overview of the relative changes in rigidity across the AL-09 V_L sequence than was able to be reported previously[9].

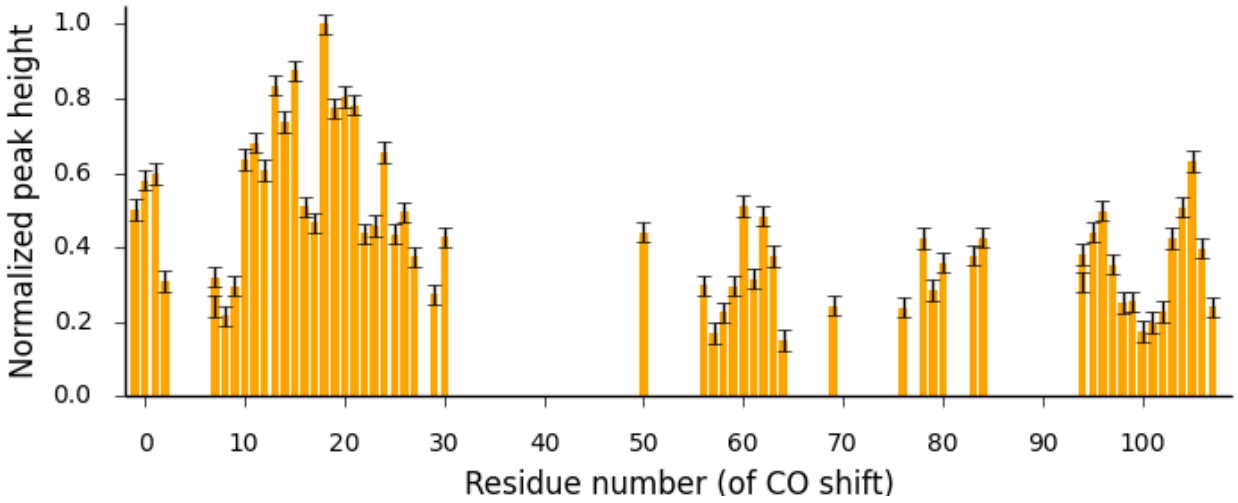

Figure 4.8. Normalized CANCOCA peak heights for assigned residues in the CANCOCA 4D spectrum with SPC-8 q64 (1.28 ms) mixing and no J-decoupling pulses. Data were acquired at 750 MHz with a 3.7% NUS schedule and MAS rate of 12.500 kHz. Values correspond to the CO-CA containing residue; error bars represent the estimated noise in the data set.

Furthermore, given the new identification of site-resolved cross peaks for multiple residues within the central portion of the AL-09 V_L sequence, we were able to obtain more quantitative information reporting on the dynamics of each observed region in the AL-09 V_L fibril core. This analysis was pursued by measuring the ^{1}H-$^{13}C\alpha$ scaled dipolar couplings of several representative residues through the acquisition and fitting of dipolar dephasing trajectories obtained from $R48_3^{18}$-symmetry ^{1}H-^{13}C β-$^{13}C\alpha$ 3D experiments[50]. We applied this to the case of V58 because it exhibits a resolvable Cβ-Cα cross peak in the CC 2D spectrum and provides an opportunity to compare the resulting order parameter with those of valine residues present in the N-terminal region of the sequence (V15 and V19). Our analysis of the corresponding dephasing trajectories for each Val residue reveals that V58 exhibits a slightly lower ^{1}H-$^{13}C\alpha$ order parameter (0.97) relative to both V15 and V19 (0.98 and 0.99, respectively; Figure 4.9A and 4.10). The differences in scaled dipolar couplings between each of the three Val residues in the AL-09 V_L sequence can be easily seen in the experimental dipolar spectra shown in Figure 4.9A, in which the V58 dipolar splitting is clearly smaller than those of V15 and V19 (simulated fits of each are included in Figure 4.10). These results thus demonstrate that the backbone of V58 is partially more dynamic than V15 and V19, supporting the existence of generally more dynamic residues near the central portion of the sequence relative to the terminal portions. Interestingly, in addition to the measured order parameters for the backbone of each valine residue, chemical shift assignments for the side-chain atoms reveal further support for the apparent increased dynamics of V58. Specifically, while residues V15 and V19 exhibit side-chain methyl chemical shifts differences of 2.2 ppm and 1.0 ppm, respectively, between the corresponding Cγ1 and Cγ2 resonances, the methyl shift difference observed for V58 is 0.4 ppm. These chemical shift differences can be observed in the Val methyl region of the ^{13}C-^{13}C 2D (Figure 4.9B), in which two distinct cross peaks are resolved for the Cβ-Cγ1 and Cβ-Cγ2 correlations of both V15 and V19, while those corresponding to V58 are severely overlapped. The existence of a relatively small methyl ^{13}C chemical shift difference for V58 may be indicative of partial rotameric averaging of the Cγ1 and Cγ2 chemical shifts about the Cα-Cβ bond, as can be the case for solvent-exposed side chains

provided that the temperature is not too low to restrict hopping between rotameric states[51,52]. Indeed, the occurrence of rotameric averaging has been observed for side-chain methyl groups (i.e., Val and Leu) in the solid state for highly mobile residues, a process which can be monitored through relative changes in dipolar-based cross-peak intensities as well as methyl chemical shift differences[28]. Accordingly, our results demonstrate that the residues in the central region of the AL-09 V_L fibril core exhibit slightly faster dynamics than those located closer to the N- and C-termini, presenting novel insight into the structural components and interactions responsible for maintaining the rigid fibril assembly.

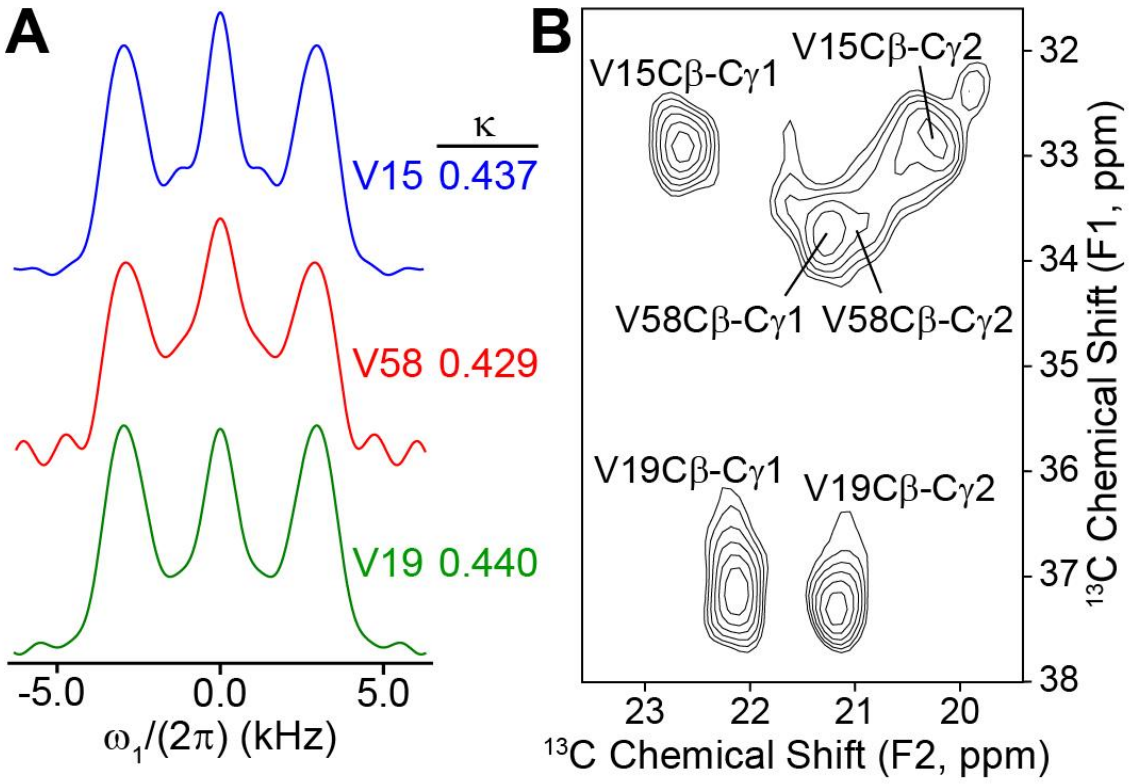

Figure 4.9. (A) ^{1}H-^{13}Ca dipolar coupling trajectories and spectra for the valine residues of AL-09 V_L fibrils, obtained from R48$_3^{18}$-symmetry ^{1}H-^{13}Cβ-^{13}Cα 3D experiments. The corresponding scaling factors obtained from fitting simulations are shown next to the dipolar spectra (simulated fits not shown). (B) Expansion of Val C β-Cγ1 and -Cγ2 region in a ^{13}C-^{13}C 2D spectrum collected with 25 ms DARR mixing at 600 MHz and 13.333 kHz MAS rate.

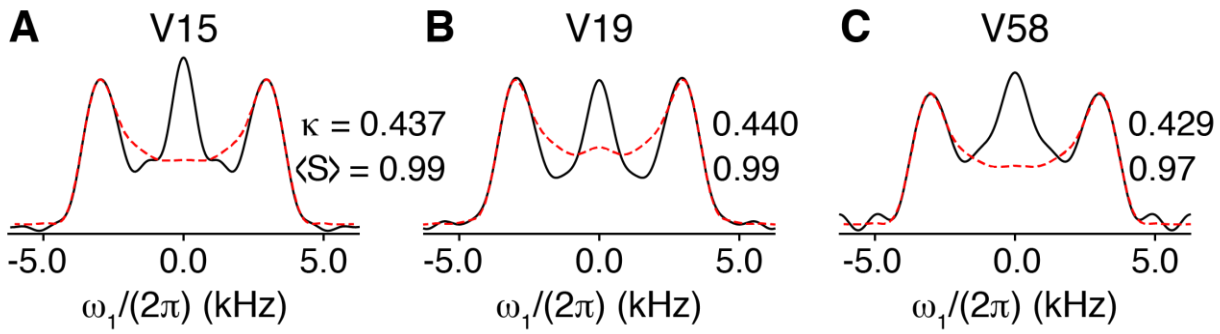

Figure 4.10. Simulated fits of the ^{1}H-^{13}Cα dipolar coupling spectra for (A) V15, (B)V19, and (C) V58, within the AL-09 V_L fibril structure (experimental, black; simulated, red). The fits reveal dipolar coupling scaling factors, κ, of 0.437, 0.440, and 0.429, which corresponds to order parameters, S, of 0.99, 0.99, and 0.97, respectively (relative to the upper order limit as observed for I21 ^{1}H-^{13}Cα, κ = 0.443; similarly, the ^{1}H-^{13}Cα of NAV exhibits a scaling factor close to this limit, κ = 0.439). Dipolar recoupling experiments were performed using the $R48_3^{18}$-symmetry sequence as the third dephasing dimension between each ^{13}C-^{13}C 2D plane.

4.5. CONCLUSIONS

In this paper, we investigated the structural and dynamical features defining the central region of the AL-09 V_L sequence within the fibril core through the implementation of J-decoupled CANCOCX and CONCACO 4D SSNMR correlation experiments, which provide enhancements in both resolution and sensitivity for spectra of uniformly [^{13}C, ^{15}N]-labeled biomolecules. These data sets enabled backbone resonance assignments to be made for more than 90% of the observable regions of the fibril structure, and subsequently site-specific analysis of the corresponding structure and dynamics for the newly identified spin-systems. Collectively, our results reveal that centrally-occurring residues of the AL-09 V_L sequence—notably, T56-T69—are involved in the core of the fibril structure, along with the highly ordered residues near the N- and C-termini that were identified previously. Interestingly, however, these central residues exhibit relatively higher flexibility and dynamics than the terminal-counterparts, as well as unique chemical shift trends suggestive of a β-turn or β-loop motif. These structural insights thus

offer several critical constraints to guide the construction of a well-supported initial model of the tertiary arrangement of the AL-09 V_L fibril core. Accordingly, we anticipate that these methods and results will inform related studies on AL disease as well as other disease-related amyloid systems.

4.6. MATERIALS AND METHODS

Preparation of U-[^{13}C,^{15}N] AL-09 V_L Fibrils. Fibril samples were prepared as described previously[30]. Following a two-week fibrillation period, AL-09 V_L fibrils were then pelleted by ultracentrifugation, dried under a steady flow of N_2 gas, packed into a 3.2 mm thin-wall SSNMR MAS rotor (Agilent Technologies, Santa Clara, CA), and brought to a 50% hydration level by mass.

SSNMR Spectroscopy. Four-dimensional MAS SSNMR CANCOCA(CX) and CONCACO(CX) experiments on uniformly [^{13}C, ^{15}N]-labeled AL-09 V_L fibrils and GB1 microcrystals were performed using a 750 MHz (^{1}H frequency) wide-bore or 500 MHz wide-bore VNMRS spectrometer (Varian, Fort Collins, CO) with a 3.2 mm Balun probe in ^{1}H-^{13}C-^{15}N mode or a 1.6 mm FastMAS probe in ^{1}H-^{13}C-^{15}N mode, respectively. All ^{13}C-detected 4D data were acquired at a MAS rate of 12.500 kHz (750 MHz) or 22.222 kHz (500 MHz) with a variable temperature (VT) set point of 0 °C. Tangent ramped cross-polarization (CP) was used to achieve heteronuclear polarization transfer from ^{1}H to ^{13}C, ^{13}C to ^{15}N, and ^{15}N to ^{13}C with contact times of 2.0 ms, 4.0 ms, and 6.0 ms, respectively[21]. SPINAL-64 ^{1}H-decoupling[24] was applied during all evolution and acquisition periods using an rf power of 65 kHz, and TPPM decoupling[26] was applied during band-selective pulse elements (as indicated in Figure 4.1)[14]. Homonuclear ^{13}C-^{13}C polarization transfer was achieved using dipolar assisted rotational resonance (DARR)[22] or supercycled POST-C_n (SPC-N)[23] mixing schemes to promote multi- or single-bond transfer, respectively. Band-selective pulses were performed on each indirect ^{13}C-dimension (ω_1 and ω_3) using r-SNOB shape elements[25] and appropriately specified pulse lengths of rotor plus a half rotor periods (τ_r), $n\tau_r + 0.5\tau_r$, for selecting the bandwidth of interest (i.e., for ^{13}Ca or ^{13}C' regions)[14]. Additional digitization and acquisition details for each 4D experiment are included in Table 4.1.

All 4D spectra were acquired according to a non-uniform sampling (NUS) regime, with sampling densities of 0.5–3.7%. Sampling schedules were generated using the nus-tool application available through the NMRBox server[53], and parameterized with the corresponding digitization limits (as specified in Table 4.1) as well as exponential biasing factors of 0.15–0.20 as inputs for all experiments.

SSNMR Data Processing. All CNCC NUS 4D data presented here were processed using the NMRPipe software package as follows[54]. Varian formatted raw data were expanded with the nusExpand.tcl script included in NMRPipe to produce a pseudo-uniform data set in which all unsampled data points are filled-in with zeroes, and converted to NMRPipe format. The directly acquired dimension was apodized with 75 Hz line-broadening and a neutral (unsquared) sine-bell window, and subsequently zero-filled prior to Fourier transformation (FT). Reconstruction of the sparsely sampled indirect dimensions was performed using the SMILE function within NMRPipe, followed by sequential apodization (60 Hz line-broadening) and FT of each dimension to generate a four-dimensional frequency-domain spectrum[31]. Phasing corrections for each dimension were determined by using NMRDraw to inspect the intermediate files from the processing output[54].

SSNMR Data Analysis. Multidimensional SSNMR data were visualized and analyzed using NMRFAM-SPARKY[55] for performing chemical shift assignments and evaluating cross-peak line widths and signal-to-noise ratio values[56]. $R48_3^{18}$-symmetry dipolar dephasing trajectories were extracted using NMRPipe and analyzed with an in-house written program for performing fitting simulations of the corresponding scaled ^{1}H-^{13}C dipolar couplings values.

4.7. SUPPLEMENTAL INFORMATION

Table 4.1. Experimental details of four-dimensional NUS SSNMR data collected on [^{13}C, ^{15}N]-labeled AL-09 V_L fibrils and GB1.

Sample	**U-[^{13}C,^{15}N] AL-09 V_L fibrils**				**U-[^{13}C,^{15}N] GB1 microcrystals**	
Experiment	**CANCOCA 4D (3.7% NUS)**	**CANCOCA 4D (2.0% NUS; soft)**	**CANCOCX 4D (2.0% NUS; soft)**	**CONCACO 4D (2.0% NUS; soft)**	**CANCOCX 4D (3.7% NUS)**	**CANCOCA 4D (0.5% NUS; soft)**
Spectrometer ^{1}H frequency	750 MHz	750 MHz	750 MHz	750 MHz	750 MHz	500 MHz
Probe	3.2 mm Balun	3.2 mm Balun	3.2 mm Balun	3.2 mm Balun	3.2 mm Balun	1.6 mm FastMAS
MAS (kHz)	12.500	12.500	12.500	12.500	12.500	22.222
Temperature set-point (ºC)	0	0	0	0	0	0
Number total rows (uniform grid)	32768	73728	73728	73728	32768	258048
Sampling density (NUS %)	3.7	2.0	2.0	2.0	3.7	0.5
Number sampled rows (NUS grid)	1200	1472	1472	1472	1200	1290
Number of scans/row	4	4	4	4	4	4
Pulse delay (s)	2.0	2.5	2.5	2.5	2.0	2.5
Total signal averaging time (h)	118	163	197	132	68	59
Polarization transfer 1	H-CA CP	H-CA CP	H-CA CP	H-CO CP	H-CA CP	H-CA CP
Transfer time (ms)	2.000	2.000	2.000	2.000	2.000	0.720
RF power (kHz)	69(H)/68(CA)	67(H)/72(CA)	69(H)/74(CA)	62(H)/75(CO)	60(H)/59(CA)	89(H)/68(CA)
Shape	tangent	tangent	tangent	tangent	tangent	tangent
Carrier (ppm)	57.5	57.5	57.5	175.0 (CP offset)	57.5	57.5
Selective pulse ω1	n/a	r-SNOB (CA)	r-SNOB (CA)	r-SNOB (CO)	n/a	r-SNOB (CA)
Pulse width (us)	n/a	680.0	440.0	840.0	n/a	495.0
RF power (kHz)	n/a	33.0	53.7	51.7	n/a	43.6
^{1}H decoupling power (kHz)	n/a	82 (TPPM)	82 (TPPM)	82 (TPPM)	n/a	120 (TPPM)
Polarization transfer 2	CA-N CP	CA-N CP	CA-N CP	CO-N CP	CA-N CP	CA-N CP
Transfer time (ms)	4.000	4.000	4.000	4.000	4.000	6.000
RF power (kHz)	22(CA)/31(N)	22(CA)/29(N)	22(CA)/29(N)	46(CO)/30(N)	22(CA)/32(N)	13(CA)/34(N)
Shape	tangent	tangent	tangent	tangent	tangent	tangent
Carrier (ppm)	117.5	117.5	117.5	175.0 (CP offset)	117.5	117.5
Polarization transfer 3	N-CO CP	N-CO CP	N-CO CP	N-CA CP	N-CO CP	N-CO CP
Transfer time (ms)	6.000	6.000	6.000	6.000	6.000	5.000
RF power (kHz)	30(N)/48(CO)	31(N)/50(CO)	31(N)/50(CO)	31(N)/23(CA)	32(N)/48(CO)	48(N)/28(CO)
Shape	tangent	tangent	tangent	tangent	tangent	tangent
Carrier (ppm)	175.0 (CP offset)	175.0 (CP offset)	175.0 (CP offset)	57.5	175.0 (CP offset)	175.0 (CP offset)
Selective pulse ω3	n/a	r-SNOB (CO)	r-SNOB (CO)	r-SNOB (CA)	n/a	r-SNOB (CO)
Pulse width (us)	n/a	840	840	360	n/a	495
RF power (kHz)	n/a	33.0	33.0	31.7	n/a	48.0
^{1}H decoupling power (kHz)	n/a	82 (TPPM)	82 (TPPM)	82 (TPPM)	n/a	120 (TPPM)

Table 4.1. (cont.)

Polarization transfer 4	SPC-8 (q64)	SPC-8 (q64)	DARR	SPC-8 (q64)	DARR	SPC-5 (q60)
Transfer time (ms)	1.28	1.28	50.00	1.28	50.00	1.08
RF power (kHz)	93.7 (C)	97.4 (C)	14.1 (H)	90.2 (C)	13.6 (H)	126.4 (C)
Increments in ω1	32	48	48	32	32	48
Sweep width (kHz)	6.250	6.250	6.250	3.125	6.250	5.556
Evol. time (ms)	5.12	7.68	7.68	10.24	5.12	8.64
Increments in ω2	64	48	48	48	64	84
Sweep width (kHz)	6.250	3.125	3.125	3.125	6.250	5.556
Evol. time (ms)	10.24	15.36	15.36	15.36	10.24	15.12
Increments in ω3	16	32	32	48	16	64
Sweep width (kHz)	3.125	3.125	3.125	6.25	3.125	5.556
Evol. or Acq. time (ms)	5.12	10.24	10.24	7.68	5.12	11.52
Increments in ω4	2000	2000	2000	2000	2000	2000
Sweep width (kHz)	100.000	100.000	100.000	100.000	100.000	100.000
Acq. time (ms)	20.00	20.00	20.00	20.00	20.00	20.00
Decoupling	SPINAL64	SPINAL64	SPINAL64	SPINAL64	SPINAL64	SPINAL64
RF power (kHz)	64.6	64.6	64.6	64.6	62.2	101.8

4.8. REFERENCES

(1) Blancas-Mejía, L. M.; Ramirez-Alvarado, M. Systemic amyloidoses. *Annu. Rev. Biochem.* **2013**, *82*, 745–774.

(2) Baden, E. M.; Sikkink, L. A.; Ramirez-Alvarado, M. Light chain amyloidosis – current findings and future prospects. *Curr. Protein Pept. Sci.* **2009**, *10*, 500–508.

(3) Mankad, A. K.; Sesay, I.; Shah, K. B. Light-chain cardiac amyloidosis. *Curr. Probl. Cancer* **2017**, *41*, 144–156.

(4) Grogan, M.; Dispenzieri, A.; Gertz, M. A. Light-chain cardiac amyloidosis: strategies to promote early diagnosis and cardiac response. *Heart* **2017**, *103*, 1065–1072.

(5) Velander, P.; Wu, L.; Henderson, F.; Zhang, S.; Bevan, D. R.; Xu, B. Natural product-based amyloid inhibitors. *Biochem. Pharmacol.* **2017**, *139*, 40–55.

(6) Nuvolone, M.; Merlini, G. Systemic amyloidosis: novel therapies and role of biomarkers. *Nephrol. Dial. Transplant.* **2017**, *32*, 770–780.

(7) Gertz, M. A.; Landau, H.; Comenzo, R. L.; Seldin, D.; Weiss, B.; Zonder, J.; Merlini, G.; Schonland, S.; Walling, J.; Kinney, G. G.; Koller, M.; Schenk, D. B.; Guthrie, S. D.; Liedtke, M. First-in-Human Phase I/II Study of NEOD001 in Patients With Light Chain Amyloidosis and Persistent Organ Dysfunction. *J. Clin. Oncol.* **2016**, *34*, 1097–1103.

(8) Jiang, L.; Liu, C.; Leibly, D.; Landau, M.; Zhao, M. L.; Hughes, M. P.; Eisenberg, D. S. Structure-based discovery of fiber-binding compounds that reduce the cytotoxicity of amyloid beta. *eLife* **2013**, *2*, e00857.

(9) Piehl, D. W.; Blancas-Mejia, L. M.; Wall, J. S.; Kennel, S. J.; Ramirez-Alvarado, M.; Rienstra, C. M. Immunoglobulin light chains form an extensive and highly ordered fibril involving the N- and C-termini. *ACS Omega* **2017**, *2*, 712–720.

(10) Hora, M.; Sarkar, R.; Morris, V.; Xue, K.; Prade, E.; Harding, E.; Buchner, J.; Reif, B. MAK33 antibody light chain amyloid fibrils are similar to oligomeric precursors. *PLoS One* **2017**, *12*, e0181799.

(11) Franks, W. T.; Kloepper, K. D.; Wylie, B. J.; Rienstra, C. M. Four-dimensional heteronuclear correlation experiments for chemical shift assignment of solid proteins. *J. Biomol. NMR* **2007**, *39*, 107–131.

(12) Wylie, B. J.; Bhate, M. P.; McDermott, A. E. Transmembrane allosteric coupling of the gates in a potassium channel. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 185–190.

(13) Straus, S. K.; Bremi, T.; Ernst, R. R. Resolution enhancement by homonuclear J decoupling in solid-state MAS NMR. *Chem. Phys. Lett.* **1996**, *262*, 709–715.

(14) Li, Y.; Wylie, B. J.; Rienstra, C. M. Selective refocusing pulses in magic-angle spinning NMR: characterization and applications to multi-dimensional protein spectroscopy. *J. Magn. Reson.* **2006**, *179*, 206–216.

(15) Higman, V. A.; Flinders, J.; Hiller, M.; Jehle, S.; Markovic, S.; Fiedler, S.; van Rossum, B. J.; Oschkinat, H. Assigning large proteins in the solid state: a MAS NMR resonance assignment strategy using selectively and extensively 13C-labelled proteins. *J. Biomol. NMR* **2009**, *44*, 245–260.

(16) Hoch, J. C.; Maciejewski, M. W.; Mobli, M.; Schuyler, A. D.; Stern, A. S. Nonuniform sampling and maximum entropy reconstruction in multidimensional NMR. *Acc. Chem. Res.* **2014**, *47*, 708–717.

(17) Palmer, M. R.; Suiter, C. L.; Henry, G. E.; Rovnyak, J.; Hoch, J. C.; Polenova, T.; Rovnyak, D. Sensitivity of nonuniform sampling NMR. *J. Phys. Chem. B* **2015**, *119*, 6502–6515.

(18) Linser, R.; Bardiaux, B.; Andreas, L. B.; Hyberts, S. G.; Morris, V. K.; Pintacuda, G.; Sunde, M.; Kwan, A. H.; Wagner, G. Solid-state NMR structure determination from diagonal-compensated, sparsely nonuniform-sampled 4D proton-proton restraints. *J. Am. Chem. Soc.* **2014**, *136*, 11002–11010.

(19) Xiang, S.; Chevelkov, V.; Becker, S.; Lange, A. Towards automatic protein backbone assignment using proton-detected 4D solid-state NMR data. *J. Biomol. NMR* **2014**, *60*, 85–90.

(20) Sergeyev, I. V.; Itin, B.; Rogawski, R.; Day, L. A.; McDermott, A. E. Efficient assignment and NMR analysis of an intact virus using sequential side-chain correlations and DNP sensitization. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, 5171–5176.

(21) Baldus, M.; Petkova, A. T.; Herzfeld, J.; Griffin, R. G. Cross polarization in the tilted frame: assignment and spectral simplification in heteronuclear spin systems. *Mol. Phys.* **1998**, *95*, 1197–1207.

(22) Takegoshi, K.; Nakamura, S.; Terao, T. 13C–1H dipolar-assisted rotational resonance in magic-angle spinning NMR. *Chem. Phys. Lett.* **2001**, *344*, 631–637.

(23) Hohwy, M.; Rienstra, C. M.; Jaroniec, C. P.; Griffin, R. G. Fivefold symmetric homonuclear dipolar recoupling in rotating solids: Application to double quantum spectroscopy. *J. Chem. Phys.* **1999**, *110*, 7983–7992.

(24) Comellas, G.; Lopez, J. J.; Nieuwkoop, A. J.; Lemkau, L. R.; Rienstra, C. M. Straightforward, effective calibration of SPINAL-64 decoupling results in the enhancement of sensitivity and resolution of biomolecular solid-state NMR. *J. Magn. Reson.* **2011**, *209*, 131–135.

(25) Kupce, E.; Boyd, J.; Campbell, I. D. Short selective pulses for biochemical applications. *J. Magn. Reson. B* **1995**, *106*, 300–303.

(26) Bennett, A. E.; Rienstra, C. M.; Auger, M.; Lakshmi, K. V.; Griffin, R. G. Heteronuclear decoupling in rotating solids. *J. Chem. Phys.* **1995**, *103*, 6951–6958.

(27) Paramasivam, S.; Suiter, C. L.; Hou, G.; Sun, S.; Palmer, M.; Hoch, J. C.; Rovnyak, D.; Polenova, T. Enhanced sensitivity by nonuniform sampling enables multidimensional MAS NMR spectroscopy of protein assemblies. *J Phys Chem B* **2012**, *116*, 7416-7427.

(28) Franks, W. T.; Zhou, D. H.; Wylie, B. J.; Money, B. G.; Graesser, D. T.; Frericks, H. L.; Sahota, G.; Rienstra, C. M. Magic-angle spinning solid-state NMR spectroscopy of the β1

immunoglobulin binding domain of protein G (GB1): 15N and 13C chemical shift assignments and conformational analysis. *J. Am. Chem. Soc.* **2005**, *127*, 12291–12305.

(29) Zhou, D. H. H.; Kloepper, K. D.; Winter, K. A.; Rienstra, C. M. Band-selective 13C homonuclear 3D spectroscopy for solid proteins at high field with rotor-synchronized soft pulses. *J. Biomol. NMR* **2006**, *34*, 245–257.

(30) Piehl, D. W.; Blancas-Mejia, L. M.; Ramirez-Alvarado, M.; Rienstra, C. M. Solid-state NMR chemical shift assignments for AL-09 VL immunoglobulin light chain fibrils. *Biomol. NMR Assign.* **2017**, *11*, 45–50.

(31) Ying, J.; Delaglio, F.; Torchia, D. A.; Bax, A. Sparse multidimensional iterative lineshape-enhanced (SMILE) reconstruction of both non-uniformly sampled and conventional NMR data. *J. Biomol. NMR* **2017**, *68*, 101–118.

(32) Zhang, H. Y.; Neal, S.; Wishart, D. S. RefDB: A database of uniformly referenced protein chemical shifts. *J. Biomol. NMR* **2003**, *25*, 173–195.

(33) Spera, S.; Bax, A. Empirical correlation between protein backbone conformation and Cα and Cβ 13C nuclear magnetic resonance chemical shifts. *J. Am. Chem. Soc.* **1991**, *113*, 5490–5492.

(34) Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; Nakatani, E.; Schulte, C. F.; Tolmie, D. E.; Kent Wenger, R.; Yao, H.; Markley, J. L. BioMagResBank. *Nucleic Acids Res.* **2008**, *36*, D402–408.

(35) Shen, Y.; Bax, A. Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J. Biomol. NMR* **2013**, *56*, 227–241.

(36) Wasmer, C.; Lange, A.; Van Melckebeke, H.; Siemer, A. B.; Riek, R.; Meier, B. H. Amyloid fibrils of the HET-s(218–289) prion form a beta solenoid with a triangular hydrophobic core. *Science* **2008**, *319*, 1523–1526.

(37) Tuttle, M. D.; Comellas, G.; Nieuwkoop, A. J.; Covell, D. J.; Berthold, D. A.; Kloepper, K. D.; Courtney, J. M.; Kim, J. K.; Barclay, A. M.; Kendall, A.; Wan, W.; Stubbs, G.; Schwieters, C. D.; Lee, V. M. Y.; George, J. M.; Rienstra, C. M. Solid-state NMR structure of a pathogenic fibril of full-length human α-synuclein. *Nat. Struct. Mol. Biol.* **2016**, *23*, 409–415.

(38) Murray, D. T.; Kato, M.; Lin, Y.; Thurber, K. R.; Hung, I.; McKnight, S. L.; Tycko, R. Structure of FUS protein fibrils and its relevance to self-assembly and phase separation of low-complexity domains. *Cell* **2017**, *171*, 615–627 e616.

(39) Helmus, J. J.; Surewicz, K.; Nadaud, P. S.; Surewicz, W. K.; Jaroniec, C. P. Molecular conformation and dynamics of the Y145Stop variant of human prion protein. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 6284–6289.

(40) Baden, E. M.; Owen, B. A.; Peterson, F. C.; Volkman, B. F.; Ramirez-Alvarado, M.; Thompson, J. R. Altered dimer interface decreases stability in an amyloidogenic protein. *J. Biol. Chem.* **2008**, *283*, 15853–15860.

(41) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res* **2000**, *28*, 235–242.

(42) Fox, J. C.; Tyler, R. C.; Guzzo, C.; Tuinstra, R. L.; Peterson, F. C.; Lusso, P.; Volkman, B. F. Engineering Metamorphic Chemokine Lymphotactin/XCL1 into the GAG-Binding, HIV-Inhibitory Dimer Conformation. *ACS Chem. Biol.* **2015**, *10*, 2580–2588.

(43) Vernet, R. D.; Boekelheide, V. Nuclear magnetic resonance spectroscopy. Ring-current effects on carbon-13 chemical shifts. *Proc. Natl. Acad. Sci. U.S.A.* **1974**, *71*, 2961–2964.

(44) Blanchard, L.; Hunter, C. N.; Williamson, M. P. The effect of ring currents on carbon chemical shifts in cytochromes. *J. Biomol. NMR* **1997**, *9*, 389–395.

(45) de Dios, A. C.; Pearson, J. G.; Oldfield, E. Secondary and Tertiary Structural Effects on Protein NMR Chemical Shifts: An ab Initio Approach. *Science* **1993**, *260*, 1491–1496.

(46) Asakura, T.; Ashida, J.; Yamane, T.; Kameda, T.; Nakazawa, Y.; Ohgo, K.; Komatsu, K. A repeated beta-turn structure in poly(Ala-Gly) as a model for silk I of Bombyx mori silk fibroin studied with two-dimensional spin-diffusion NMR under off magic angle spinning and rotational echo double resonance. *J. Mol. Biol.* **2001**, *306*, 291–305.

(47) Saito, H.; Tabeta, R.; Asakura, T.; Iwanaga, Y.; Shoji, A.; Ozaki, T.; Ando, I. High-Resolution 13C NMR-Study of Silk Fibroin in the Solid-State by the Cross-Polarization Magic Angle Spinning Method - Conformational Characterization of Silk-I and Silk-Ii Type Forms of Bombyx-Mori Fibroin by the Conformation-Dependent 13C Chemical-Shifts. *Macromolecules* **1984**, *17*, 1405–1412.

(48) Asakura, T.; Ohgo, K.; Komatsu, K.; Kanenari, M.; Okuyama, K. Refinement of repeated beta-turn structure for silk I conformation of Bombyx mori silk fibroin using 13C solid-state NMR and X-ray diffraction methods. *Macromolecules* **2005**, *38*, 7397–7403.

(49) Tycko, R. Physical and structural basis for polymorphism in amyloid fibrils. *Protein Sci.* **2014**, *23*, 1528–1539.

(50) Zhao, X.; Eden, M.; Levitt, M. H. Recoupling of heteronuclear dipolar interactions in solid-state NMR using symmetry-based pulse sequences. *Chem. Phys. Lett.* **2001**, *342*, 353–361.

(51) Hong, M.; Mishanina, T. V.; Cady, S. D. Accurate measurement of methyl 13C chemical shifts by solid-state NMR for the determination of protein side chain conformation: the influenza a M2 transmembrane peptide as an example. *J Am Chem Soc* **2009**, *131*, 7806–7816.

(52) Chou, J. J.; Case, D. A.; Bax, A. Insights into the mobility of methyl-bearing side chains in proteins from (3)J(CC) and (3)J(CN) couplings. *J Am Chem Soc* **2003**, *125*, 8959–8966.

(53) Maciejewski, M. W.; Schuyler, A. D.; Gryk, M. R.; Moraru, II; Romero, P. R.; Ulrich, E. L.; Eghbalnia, H. R.; Livny, M.; Delaglio, F.; Hoch, J. C. NMRbox: A Resource for Biomolecular NMR Computation. *Biophys. J.* **2017**, *112*, 1529–1534.

(54) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. NMRPipe: a multidimensional spectral processing system based on unix pipes. *J. Biomol. NMR* **1995**, *6*, 277–293.

(55) Lee, W.; Tonelli, M.; Markley, J. L. NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* **2015**, *31*, 1325–1327.

(56) Goddard, T. D.; Kneller, D. G. SPARKY 3. University of California, San Francisco, **2008**.

CHAPTER 5: Towards a Robust and Streamlined Strategy for Assigning Chemical Shifts in Proteins Using Peak Intensity and Overlap-Weighted Scoring on NUS SSNMR Data

5.1. NOTES AND ACKNOWLEDGEMENTS

Dennis W. Piehl, Michael J. Hallock, Manali Ghosh, Adedolapo M. Ojoawo, Xiangyan Shi, Chad M. Rienstra. This research is supported by the University of Illinois (Centennial Scholars Award to C.M.R.). D.W.P. was an American Heart Association Predoctoral Fellow (15PRE25100008).

5.2. ABSTRACT

We introduce a robust method for the computationally-aided assignment of protein chemical shifts in multidimensional solid-state nuclear magnetic resonance (SSNMR) data that combines peak intensity and frequency overlap information into a separate objective within the scoring function of the multi-objective search for optimal assignment solutions, and show that this offers significant improvement over currently available algorithms. The additional peak intensity and overlap score (PIOS) objective is implemented as a modification to the non-dominated sorting genetic algorithm/Monte Carlo programs presented previously which we refer to as PIOS/NSGA-II/MC, and alleviates several major limitations of the original programs, most notably the tolerance for assigning non-manually analyzed experimental data. In particular, rather than restricting the judgment to solely “good” or “bad” as previously was the case, we demonstrate how the PIOS objective provides an experimentally-based measure of the “goodness” of a given possible assignment, through the inclusion of cross-peak signal-to-noise ratios. In this manner, the presented program is more readily applicable to realistic SSNMR data sets, which may contain a number of “data imperfections” such as noise, artifacts, spin-system degeneracy (and thus assignment ambiguity), as well as the presence of long or short-range inter-residue cross-peaks in a particular spin system. Finally, through the combination of the non-uniform sampling and data reconstruction methods with the SMILE algorithm, we describe (and provide the computational

tools for) a more streamlined strategy for the analysis of 3D and 4D SSNMR spectra and through-put into the PIOS/NSGA-II/MC chemical shift assignment program.

5.3. INTRODUCTION

The determination of a protein structure by magic-angle spinning (MAS) solid-state nuclear magnetic resonance (SSNMR) spectroscopy generally begins with the acquisition of a variety of multidimensional homo- and heteronuclear correlation spectra to obtain the ^{13}C and ^{15}N chemical shift information of the biomolecule of interest[1]. In addition to providing initial insight on the secondary structure of the protein through correlation of the secondary chemical shifts with empirical trends[2,3], having knowledge of the specific atoms to which various cross peaks correspond can enable a qualitative assessment of dynamics based on differences in cross-peak intensity resulting from dipolar coupling-based polarization transfer experiments. Most importantly, these chemical shift assignments are necessary for performing subsequent structural and dynamics measurements on the molecule, including the determination of internuclear distance restraints[4,5], backbone order parameters[6], side-chain motions[7], as well as the investigation of how chemical shift perturbations relate to specific biomolecular activities or interactions[8]. Ultimately, these pieces of information can be combined into simulated annealing calculations for obtaining a final three-dimensional protein structure[9].

The process of assigning chemical shifts in protein SSNMR data typically involves (at minimum) the acquisition and analysis of several ^{15}N and ^{13}C 3D correlation experiments, including a $^{15}N_i$-$^{13}C\alpha_i$-$^{13}CX_i$ (NCACX), a $^{15}N_i$-$^{13}C'_{i-1}$-$^{13}CX_{i-1}$ (NCOCX), and a $^{13}C\alpha_i$-$^{15}N_i$-($^{13}C'_{i-1}$)-$^{13}CX_{i-1}$ (CAN(co)CX), which together provide the backbone connectivity information necessary for enabling the step-wise assignment of chemical shifts for each residue along the protein sequence[10,11]. This general assignment strategy is known as a "backbone-walk[1]." The collection of four-dimensional heteronuclear correlation spectra can provide an additional set of chemical shift information by further removing chemical shift degeneracy to resolve cross peaks that may be overlapped in 3D experiments[12], as well as offer another way of obtaining

a set of distance restraints for structure determination[13]. As such, this strategy is particularly useful for large or complex biomolecular systems that may exhibit substantial signal overlap or degeneracy, and is becoming routinely used in SSNMR studies with the availability of more efficient data acquisition and processing methods such as non-uniform sampling (NUS)[14,15] combined with the use of the reconstruction algorithm, SMILE[16].

While these established methods collectively provide a generalized and robust strategy for the assignment of chemical shifts in solid biomolecules, due to the amount of time and manual effort required to perform each of these tasks—in addition to the inherent challenges that are unique to the specific protein under study—this process can often be the most rate-limiting step in a structural pursuit. Challenges at this stage may arise due to a variety of experimental and technical factors, including extensive chemical shift degeneracy and peak overlap, low inherent sensitivity, the existence of multiple protein conformations within the sample, the absence of cross-peak intensity for highly dynamic regions along the backbone (and thus an impediment to the backbone walk procedure), as well as the practical limits of analyzing highly complex multidimensional data by manual means. These complications naturally become further exacerbated with increasing size of the molecule, as this leads to a higher extent of spectral congestion as well as greater uncertainty with respect to the unique spin system to which a peak corresponds. As such, despite having the technical advantage and potential to study insoluble proteins of any size, the feasibility and efficiency of assigning chemical shifts to such large systems still remains a major pressing factor in SSNMR studies.

A tremendous amount of research effort has focused on addressing this challenge by improving the computational tools and software that are currently available for assigning chemical shifts of protein SSSNMR data. In addition to the development and active support of freely-available software packages for visualizing and analyzing multidimensional NMR data[17,18], concerted efforts have been directed toward design of fast and accurate programs for computationally aiding the assignment process of protein chemical shifts. In particular, MCASSIGN[19] and its successor, MCASSIGN2[20], introduced by Tycko et al. and Hu et al., respectively, uses a Monte Carlo and simulated annealing (MC/SA) approach to the

assignment problem by randomly sampling the combination of spin systems from a set of given lists (e.g., an organized list of peak groups from a NCACX or NCOCX spectrum, in which all peaks sharing the same ^{15}N and $^{13}C\alpha$ or $^{13}C'$ frequencies are listed together), and assesses a specific set of possible connections (specified by the user) between the two spin systems to determine if the connection is "good," "bad," or falls on the boundary of an assignment ("edge"). These values are totaled for each given combination of spin systems (i.e., an assignment solution) and are combined with increasing weight upon each full program iteration (the simulated annealing component) to give a final score that is used to determine if the assignment should be kept or rejected. Indeed, this program offers a straightforward way of sampling all of the possible connections of a given list of spin systems to eventually arrive at a solution with the most number of "used" and "good" connections but fewest number of penalties ("bad" and "edge" connections).

The MCASSIGN2 method was later adopted by Yang et al.[21], in which they introduced a non-dominated sorting genetic algorithm (NSGA-II)[22] into the MC/SA program as a way to fully harness the multi-objective aspect of the assignment method and reduce the time spent on sampling redundant assignment solutions. Through the function of the NSGA-II routine, the program sorts each group of assignment solutions (i.e. a "generation") based on how well each individual dominates the others with respect to each objective (i.e., "good," "bad," "edge," and "used"), and retains only the best solutions to generate an "offspring" population that inherits a subset of the assignments from the parent generation. This continuous process of sorting the group of individuals by their objectives ultimately produces a set of rankings in which those solutions which dominate in at least one objective are proscribed a "Pareto order" of 1, and are otherwise known as Pareto-frontier solutions. Thus, rather than relying solely on the random sampling and scoring of assignment solutions by a Monte Carlo approach, the addition of this genetic algorithm allows for optimal solutions to be identified in a much more directed and efficient manner. Further, although one general concern for the use of such an algorithm is the high tendency to drive the solutions into a local minimum, the author's decision to integrate this with the MC/SA algorithm

offers a powerful way to obtain the optimal solution while still allowing for enough random sampling potentially escape these wells[21].

While the MCASSIGN2 and combined NSGA-II/MC programs each establish an adaptable and highly functional approach to the assignment of protein chemical shifts given any variety of multidimensional SSNMR spectra, they both still suffer from a number of shortcomings that can lead to severe errors in the assignment simulation. One of these limitations is the "all-or-nothing" manner by which the connection between a given set of spin systems is declared as either "good" or "bad", based on the occurrence of a single matching or non-matching of frequencies. As a result, even solutions that exhibit all matching connections except one (e.g., due to the presence of an additional peak in the spin system) will be penalized and may be excluded later from the list by restricted sorting steps, as both programs were designed to remove any assignment solutions with bad connections near the end of the calculation[20,21]. While such a tactic works reasonably well for properly analyzed data sets in which no mistakes are included in the preparation of input spin-system lists, the penalization of an entire spin system based on a single misaligned connection restricts the program to only being able to tolerate "error-free" data.

Recently, Murray et al. report the use of a modified form of the MCASSIGN2 program (referred to as mcassign2c), in which the quantity of "good" is scaled between entirely good and entirely bad based on the amount of separation between the two frequencies being connected[23]. We believe this modification represents an important step towards properly accounting for the quality of a "good" connection without giving it an equal score as another solution that may exhibit slightly closer spin-system alignment. This is especially important for instances of spin-system degeneracy, as two residues exhibiting very similar chemical shifts in two or three dimensions may both be considered equally good based on the original "all-or-nothing" approach, even if one of the spin systems matches slightly better than the other. However, the continued treatment of "bad" connections as a penalty to the assignment process still leaves open the possibility of excluding possibly valid assignments that contain a small number of non-matching

connections (e.g., due to the additional inclusion of an artifact peak, inter-residue correlation, or degenerate spin-system peak within a single spin-system).

Here, we present a highly robust adaptation of the previous programs which combines the peak intensity and frequency overlap of a given spin-system comparison as a way of scoring the quality of the connection. In order to build upon the MC and genetic algorithm functionalities offered by the previous programs, we implemented this peak intensity and overlap score (PIOS) as a new objective within the combined PIOS-weighted NSGA-II/MC program. We apply this method to the β1 immunoglobulin binding domain of protein G (GB1) as a test case, and demonstrate the improvement that it offers relative to the previous programs, which becomes most revealing for cases involving imperfect and unanalyzed data sets. Furthermore, we describe a strategy for streamlining the workflow between data acquisition by NUS and protein chemical shift assignments by PIOS/NSGA-II/MC, in which we string together several SSNMR methods and tools through the use of Python scripting, and we demonstrate the success of this automated process for fibril form of the amyloidogenic light-chain protein, AL-09 V_L.

5.4. METHODS

5.4.1. Overview of modified PIOS-weighted NSGA-II/MC algorithm.

The PIOS-weighted NSGA-II/MC assignment program that we present here for computationally-aiding the chemical shift assignment process is based on the NSGA-II/MC program previously introduced by Yang et al.[21], which combines the MCASSIGN2 algorithm[20] with a multi-objective genetic algorithm for more efficient searching of optimal assignment solutions. As such, the general flow of logic and program structure remain very similar, but with the addition of several modifications at the scoring step of the algorithm as well as the treatment of bad assignments, which we describe in the following section and Supplemental Materials 5.7.1. The source code is written in Fortran 95 (provided upon request) and may readily be compiled on Linux, OSX, or Windows operating systems using the freely-available PGI Community Edition compiler.

A flowchart illustrating each of the key steps in the program is provided in Figure 5.1. The initial input files include: (1) the amino acid sequence of the protein (in one-letter code); (2) a list of spin systems obtained from each spectrum being used, containing the chemical shift frequencies and corresponding SNR values for each peak in the spin system as well as a header line specifying the estimated uncertainty in peak position of each atom type for all spin systems; (3) a connection file specifying the connectivity and relation of each frequency within one spin-system list with another; and last (4) a control file that points to each of the input and desired output file locations and contains the list of user-specifiable run options, including the number of individuals to form the initial group of assignments (and thus the number of resulting solutions from the calculation), the number of full program steps as well as the number of MC and NSGA-II attempts to perform at each iteration, the probability of the program forming a null mutation, additional NSGA-II-specific settings such as the gene pool size as well as crossover and mutation rates for controlling the diversity of offspring from the parent generation, and finally MC/SA-related weighting factor value limits to use for each objective (w_1, "good"; w_2, "bad"; w_3, "edge"; w_4, "used"; and w_5, PIOS) in calculating the score of the individual assignment, which are incremented linearly from the minimum to maximum values over each full program iteration.

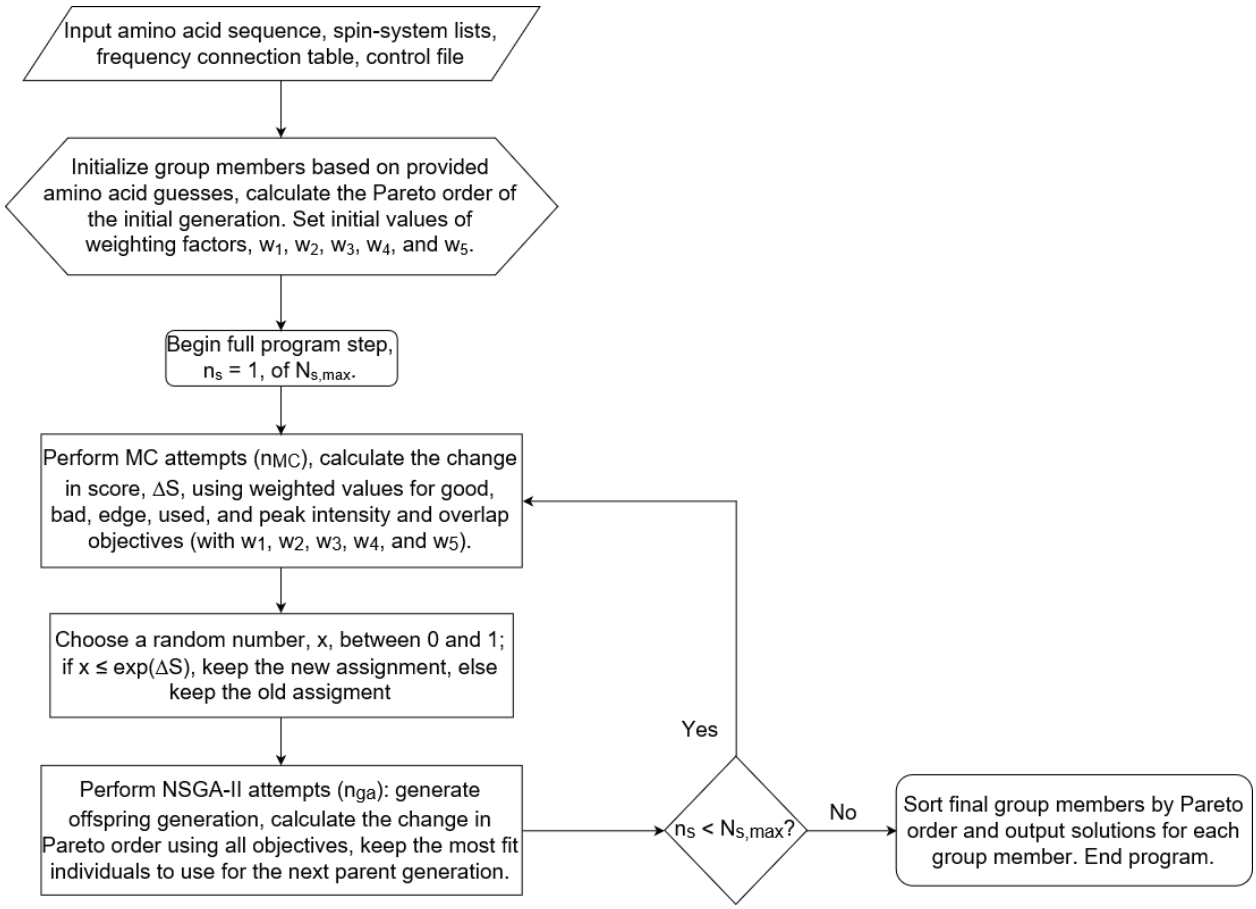

Figure 5.1. Flowchart of the control logic scheme followed by the PIOS-weighted NSGA-II/MC chemical shift assignment program.

Upon initialization of the calculation with a specified or clock time-based seed number, the program generates a starting group of individuals (i.e., complete sets of assignments) by randomly selecting spin systems from the input lists which have amino acid type guesses that are consistent with the corresponding residue in the protein sequence. If desired, it is also possible to provide the output file from a previous calculation to form the starting group individuals for a new calculation; however, the size of the group as well as the number and type of spin-system lists used must remain the same. The initial group is then evaluated for the overall fitness of the individuals by Pareto ordering, and is subsequently directed into the MC/SA procedure.

For each MC attempt, n_{mc}, specified in the control file, the algorithm begins with one of the individuals of the group and calls a mutation subroutine which randomly either changes one of the spin systems in the current set of assignments to another spin system (without violating the amino acid type or degeneracy of the given spin system) or sets it to a null spin system (based on the null probability specified at input). Within this process, all five objectives (N_{good}, N_{bad}, N_{edge}, N_{used}, and Q_{PIO}) are determined for the given assignment by looping over each possible connection of frequencies between each spin-system list for every residue in the sequence, and tabulating the number of good, bad, and edge connections that are made, the number of spin systems used, and the quality score based on the PIOS calculation described in the next section. The overall score, *S*, of the individual both before and after the mutation is calculated from the combination of the quantities for each objective, scaled by the corresponding weighting factor, *w*, as follows:

$$S = w_1 N_{\text{good}} - w_2 N_{\text{bad}} - w_3 N_{\text{edge}} + w_4 N_{\text{used}} + w_5 Q_{\text{PIO}}^{Tot,\dagger}, \quad (1)$$

where N_{good}, N_{bad}, and N_{edge}, represent the number of good, bad, and edge connections respectively, N_{used}, represents the number of used spin systems, and $Q_{\text{PIO}}^{Tot,\dagger}$ represents the total quality score based on peak intensity and overlap of all connections (†indicates that this value is divided by 100 for this scoring calculation, because it is scaled up by 100 when tabulated to keep it in integer form). After a new assignment has been made (e.g., through the change of a single spin-system assignment for a random residue in the sequence), the change in score is calculated as the difference of the two independent scores from each assignment,

$$\Delta S = S_{\text{new}} - S_{\text{old}}. \quad (2)$$

Last the MC/SA component uses Metropolis' criterion to decide whether the new assignment should be retained by choosing a random number, x_{rand}, and determining if the following inequality holds true:

$$x_{\text{rand}} \leq e^{\Delta S}. \quad (3)$$

If the condition is not met, then the program discards the new assignment and reverts back to the previous assignment before cycling back for the next MC attempt. This process continues until all MC attempts,

N_{MC}, have been completed for all individuals of the group, all of which are then carried over to the NSGA-II component to use as the initial group of assignments. Upon the next round of MC attempts (i.e., after a full calculation cycle of the program has been completed, in which the number of steps, n_s, increases by 1), the weighting factors for each objective are incremented according to the specified minimum and maximum weight values and number of total steps of the program. In this manner, this gradual incrementation of objective weighting constitutes the simulated annealing aspect of the MC/SA algorithm[19].

The program then proceeds onto the NSGA-II component of the calculation for the total number of specified attempts, N_{GA}, starting with the resulting set of assignments from the MC/SA process above onto the first attempt, $n_{ga} = 1$. The NSGA-II process begins with the determination of objective values for each individual of the given group of assignments and subsequently subjects this group to a series of mutation and crossover events (described in detail by Yang et al.[21]) in order to form an initial parent generation to use for producing a set of offspring assignments. The algorithm then cycles through the new generation and performs a second series of mutations in order to produce a completely diverse set of individuals, and removes any identical assignments. The final set of parent and offspring assignments are evaluated based on their Pareto order ranking and the most fit individuals are selected to move onto the next NSGA-II attempt, which repeats until $n_{ga} = N_{GA}$. After the cycle of attempts completes, the program returns back to the MC/SA component with the final group assignments resulting from the NSGA-II component.

5.4.2. Identification of a suitable scoring function to account for peak intensity and overlap.

In the presented PIOS-weighted NSGA-II/MC program described here, we have incorporated an additional weight to the algorithm that uses the peak intensity and extent of chemical shift overlap to provide a better measure (or score) of the quality of a tested connection, hence our referral to this augmented method as PIOS-weighted NSGA-II/MC. As depicted in the flowchart shown in Figure 5.2,

the calculation of the intensity and overlap objective score begins with the comparison of a set of frequencies from two separate spin-system lists in a given connection test, in which three quantities are calculated. First, the intensity of each peak is scored as a function of SNR according to Equations 4–5, providing a value between 0 to 1 for each peak ($I_{\text{score},1}$ and $I_{\text{score},2}$). The generalized equation that we determined to most suitably account for the influence of a peak intensity on the quality of a particular connection is a sigmoidal curve in the form of a Gompertz function, as follows:

$$f(x) = ae^{-be^{-cx}}, \tag{4}$$

where the independent variable *x* represents the peak intensity, *a* and *c* are factors that define the maximum scaling and growth of the curve along the y-axis, respectively, and *b* is the displacement factor along the x-axis. Because we wanted to normalize our scoring function to 1 to simplify the calculation process, we set *a* to 1, and subsequently adjusted the values of *b* and *c* such that a SNR, *x*, of 5 provides a score close to 0.25, while a SNR of 15 provides a score close to 1. Accordingly, the final form of the implemented intensity scoring function is written:

$$I_{\text{score}} = e^{-10e^{-0.4 \cdot SNR}}. \tag{5}$$

Our rationale behind selecting these specific scaling and growth factors was to identify a function that changes most across the range of SNR values in which the uncertainty of peak position, spin-system belonging, and authenticity (i.e., whether it is noise or not) is highest. Beyond a noise threshold of 15 σ (i.e., 15 times the noise-floor), most observed peaks can be considered to be "true peaks" in moderately signal-averaged multidimensional MAS SSNMR spectra (e.g.., several hours to a day for 2D spectra, or one to several days for higher-dimensional experiments). Further, based on negative control simulations in which the SNR of non-representative peaks within a given spin system were adjusted relative to true peaks within the spin system, the negative influence of SNR on the success of assignment solutions was found to become most severe below values of 10 for truly-representative peaks (with non-representative peak SNR values set to 30) (Figure 5.11). Although we have not included the option for users to adjust

these constants through the settings in the input control file for the PIOS-weighted NSGA-II/MC program, this may be easily changed in the program code directly if desired.

Next, the extent of overlap of the two frequencies is scored based on the chemical shifts and user-specified peak position uncertainties (ε_1 and ε_2) of each frequency, normalized to $2*\varepsilon_{\min(1,2)}$ of the peak with the narrower uncertainty range to give a value between 0 and 1, O_{score}, as follows:

$$O_{\text{score}} = \frac{\text{Overlap of peaks 1 and 2 within range of } \varepsilon_1 \text{ and } \varepsilon_2}{2*\min(\varepsilon_{1,2})}. \quad (6)$$

Subsequently, all three quantities are multiplied together to yield a product that represents the PIOS (or quality score) of the tested connection:

$$Q_{\text{PIO}} = I_{\text{score, peak 1}} * I_{\text{score, peak 2}} * O_{\text{score}}, \quad (7)$$

which provides a value of Q_{PIO} between 0 and 1, which is then scaled by 100[†] and rounded up to the nearest integer in order to meet the compatibility requirements in the program code ([†]however, the final quality score is scaled back down by 100 during the overall objective scoring calculation in Equation 1). This quality score is calculated for every possible combination of frequencies specified in the connection table (N_{rows}), for every spin system in the given assignment (N_{seq}), and is tabulated for every i^{th} possible connection (N_{connect}) throughout this process to provide a final total quality score for the entire assignment given by,

$$Q_{\text{PIO}}^{Tot} = \sum_{i=1}^{N_{connect}} Q_{\text{PIO},i} * 100,$$

$$\text{where, } N_{connect} = N_{seq} * N_{spectra} * N_{rows}, \quad (8)$$

Accordingly, our method effectively provides a way of scaling the quality of a given set of connections, as opposed to assigning them as either "entirely good" or "entirely bad," but still remains separate from the tabulation of the number of good connections. In this manner, in the event that a final set of assignment solutions exhibit the same number of good connections, one may still be able to assess the quality of each via the separate PIOS objective.

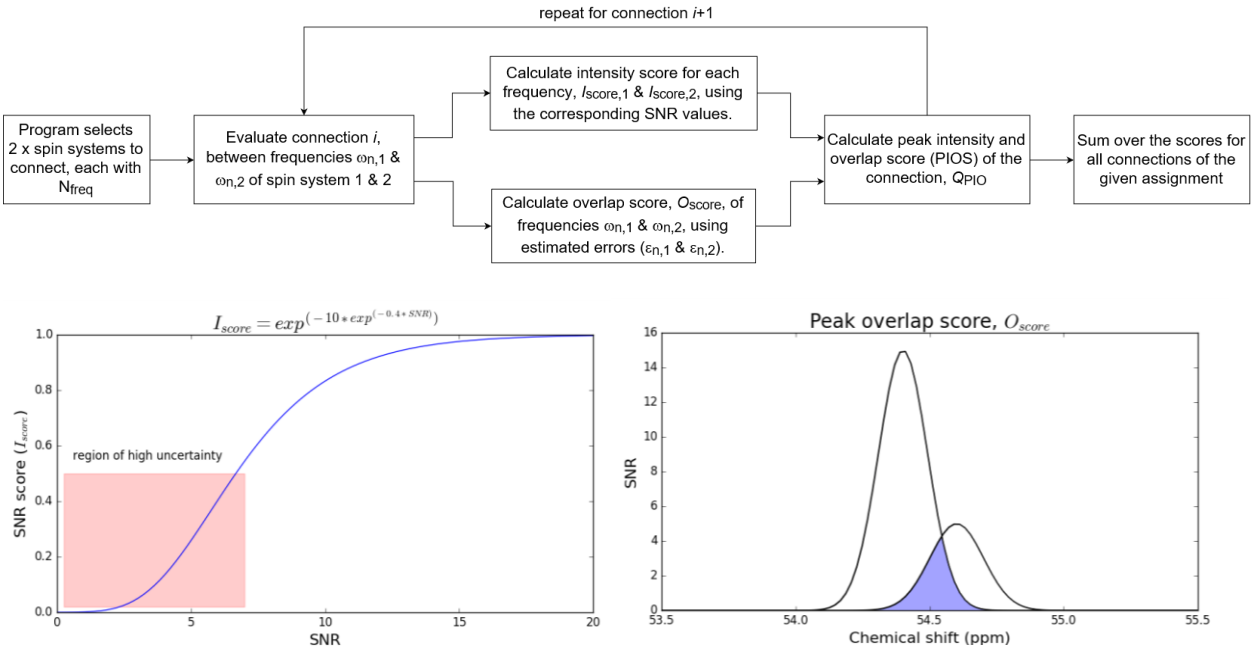

Figure 5.2. Calculation workflow for scoring the quality value of a given connection based on the SNR of each peak as well as the extent of overlap.

5.4.3. Implementation of PIOS-weighted function within the NSGA-II/MC program.

The interest in introducing the PIO-weighted score into the assignment program was to augment the robustness and performance of the previously expanded NSGA-II/MC algorithm, by incorporating valuable spectral information into an additional objective through which the program can propagate good assignments versus those that are considered bad in a more informed manner. This strategy functions in itself to help alleviate complications arising from the presence of additional peaks in a residue's list of frequencies due to spin system degeneracy or weak inter-residue correlations (e.g., > $i \pm 1$) in the SSNMR spectra through its scoring of the quality of the connection in a non-binary manner, and using peak intensity data combined with frequency alignment to inform on the extent to which the assignment "fits" (Figure 5.3). In this way, the PIOS-weighted NSGA-II/MC algorithm is highly tolerable of accepting generally less refined data sets that have not been extensively analyzed manually, since it indirectly accounts for the possible presence of a bad connection by giving the assignment a slightly lower score.

Accordingly, an important feature distinguishing this strategy from the combined NSGA-II/MC or standalone MCASSIGN2 methods is that here, final solutions that still contain a number of "bad connections" do not need to be—and should not be in many cases—completely eliminated in the assisted assignment process, which is a general requirement of the other programs[20,21]. The way in which the PIO scoring function is implemented in the program is what allows for it to do this.

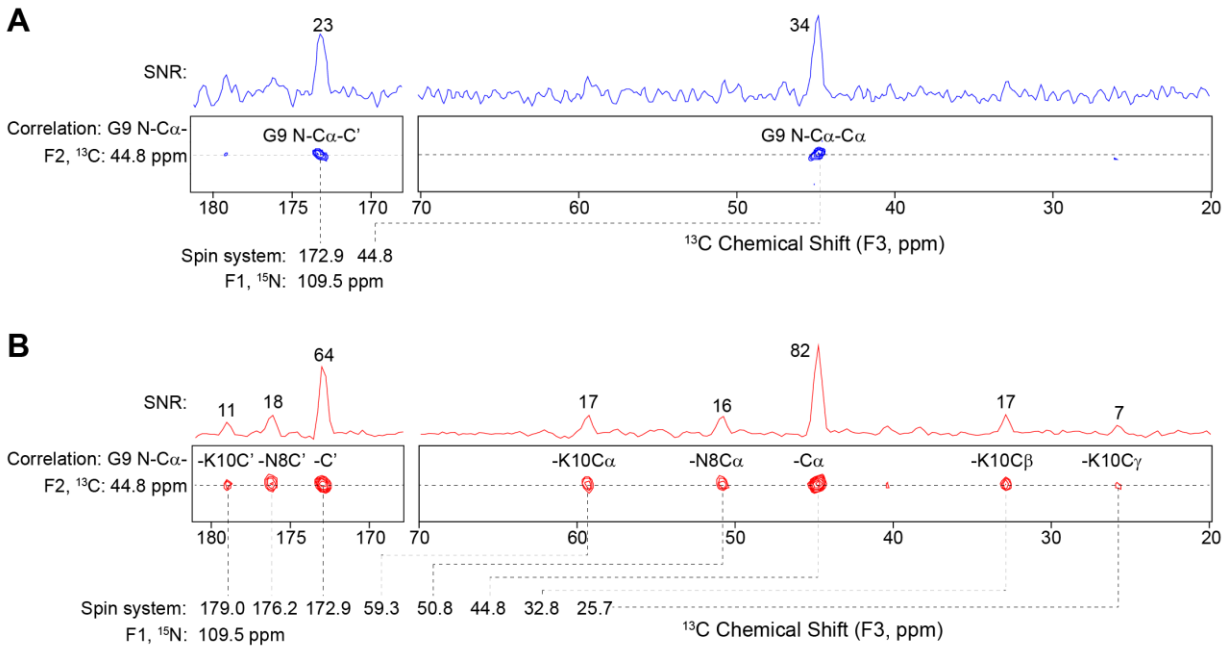

Figure 5.3. (A) 2D strip and corresponding 1D slice from a US NCACX 50 ms 3D spectrum illustrating the spin system of residue G9 in GB1 without any inter-residue correlations observed. (B) The same strip and slice of a separate NCACX 50 ms 3D (22% NUS) spectrum, which illustrates the additional presence of inter-residue peaks within the G9 spin system of GB1. Both spectra were collected with a MAS rate of 12.500 kHz at 750 MHz (^{1}H frequency) and 0 C for approximately 14 hours total.

In deciding how to incorporate the scoring function into the NSGA-II/MC algorithm, we considered and tested several options before choosing to implement it as an additional objective involved in the Pareto ordering process. An alternative route by which we also added the function was through the

direct inclusion as part of the "good" objective, as a way to simply scale the overall "goodness" of each assignment. However, because we recognized that different connections will provide different magnitudes of PIOS depending on the strength of the SNR for the given connection, and as such may suppress the retention of assignment solutions which exhibit a higher number of good connections but which may be slightly weaker or fewer. For instance, there may be a situation in which one assignment solution has only one half of the total number of possible good connections but a very high PIOS objective quantity due to the alignment of many strong peaks between multiple connections, while another solution has 75% of the total number of good connections but only a modest PIOS quantity due to the corresponding spin systems having either very low intensity or few frequencies to overlap. In such an event, the combination of the PIOS quantity into the number of good connections would result in the suppression of the second possible solution. As such, by instead implementing the PIO score as a separate objective, assignment solutions with higher good connections but lower PIO-score, or vice-versa, would both be presented as valid solutions with different maximized objectives, rather than losing one under the control of the other.

As we have emphasized, one additional benefit to adding the PIOS objective into the assignment algorithm is that the score is not penalized for making a weak connection or not making any connection at all, but rather either receives a lower quality score or no score at all, respectively. As this scaling of the score offers an indirect way of judging whether a connection is "good" or "bad", for the majority of the simulations that we run here were performed with the tabulation of bad connections turned off. Although the continued tabulation of bad connections would not be an issue for cases of near-ideal data sets—particularly for those in which the chemical shifts are listed in a specific order based on the knowledge of the specific atom type—in situations where the atom type of a particular chemical shift cannot be confidently identified, let alone whether the peak truly belongs in the given spin system, the number of tabulated bad connections can become exceedingly large and cause disruption to the sorting process even if the weights on bad are set to zero. Moreover, for nearly all of the calculations described below (which were performed using spin-system lists generated from minimally analyzed data sets) we use a connection list which specifies the possible connection between every possible ω3 frequency of one spectrum with

every possible ω3 frequency of another, with up to six ω3 frequencies present in each spin system. This was intentionally done as way to further test the robustness of the presented algorithm for instances in which the atom type to which a specific chemical shift column corresponds does not need to be known in advance of performing the assignment simulation. Yet, as a consequence of forcing the program to sample all of the possible good and bad connections between each spin system, this incidentally can result in the tabulation of a large quantity of bad (or simply, "mismatched") connections as well, which may lead to substantial complications at the sorting/ranking step further downstream in the program. Accordingly, the calculations described below avoid the tabulation of bad connections all together, allowing us to directly assess the influence of the added PIOS objective function as well as the robustness of the PIOS-weighted NSGA-II/MC program as a whole.

5.4.4. Strategies for streamlining the analysis of NUS 3D and 4D data by the PIOS-weighted NSGA-II/MC algorithm.

In the overarching interest of arriving at a fully automated approach for the assignment of chemical shifts in SSNMR spectra of proteins, we have developed and combined a collection of computational tools for streamlining the procedure involved in the translation of minimally analyzed NMR data sets into unbiased and pre-formatted spin-system lists that can be directly used as input to PIO-weighted NSGA-II/MC calculations. Our work towards this goal was aimed at addressing three major rate-limiting steps in the typical process of preparing signal lists for program input: (1) the manual picking and grouping of peaks into separate spin systems; (2) the process of determining amino acid guesses for each of the corresponding spin systems; and (3) finally organizing these spin-system lists into a readable format for the assignment algorithm. These steps previously required a substantial amount of manual effort by the user and thus not only partially defeat the purpose of establishing a robust automated assignment algorithm, but also increase the susceptibility of overall process to user-bias and error. An overview of the typical workflow and our efforts to streamline this process is illustrated in Figure 5.4.

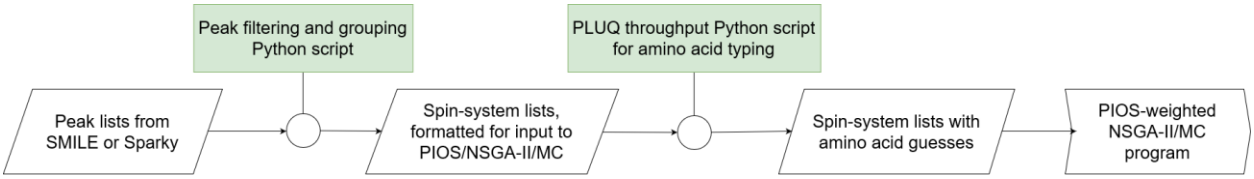

Figure 5.4. Streamlined workflow for the generation of input files for PIOS/NSGA-II/MC using raw peak lists generated from either the reconstruction of spectra using SMILE or peak picking in Sparky.

The first time-consuming step we address involves the picking and grouping of peaks in 3D and/or 4D spectra into separate spin-system lists. Typically, this requires the user to manually work through each spectrum, pick the observed cross peaks at different noise-threshold cutoffs, and then group these peaks into unassigned spin systems. While for some non-congested or highly resolved 3D SSNMR spectra this step may be done fairly effortlessly, it becomes a much greater difficulty in cases where there is significant peak overlap, spin-system degeneracy, or the presence of additional cross-peaks within a particular spin system arising from inter-residue correlations, as this can obscure what peaks truly belong to the same spin system and thus require a number of subjective decisions from the user (Figure 5.3). Indeed, several available NMR data analysis packages offer tools for expediting this process, such as through the use of automated peak-picking extensions and assignment guessing options, such as in Sparky[17,18], CcpNmr[24], NMRView[25], and NMRDraw[26]. Further, some of these programs already offer means of automatically assigning solution NMR data through the integration of other assignment software, such as the PINE plugin[27] within the NMRFAM-SPARKY package[18]. However, a generalized solution for the generation spin-system lists from peak lists for SSNMR data and the corresponding automation of their assignment remains largely limited.

To address this limitation, we have assembled an in-house script to process the output from Sparky, group the peaks into spin systems based on user-defined tolerances and convert them into a compatible format for the PIOS-weighted NSGA-II/MC program, as illustrated by the work flow in

Figure 5.5. Additionally, to further extend this utility to the case of NUS 3D and 4D SSNMR spectra, we have also incorporated the capability of the script to directly read in the log file output from SMILE[16] reconstructed NUS spectra (as made available through the NMRPipe software package)[26], which contains a list of chemical shift and data height information for all of the reconstructed peaks in the corresponding spectrum. In this manner, we are able to bypass the manual interaction with the fully processed multidimensional data set entirely (although it remains critical that the spectrum is processed correctly, including phasing specifications in the SMILE function). Together, the immense reduction in time required to acquire 3D and 4D SSNMR data by means of NUS, reconstruct the spectrum and immediately obtain a list of spin-systems that can be incorporated into the PIOS-weighted NSGA-II/MC assignment program offers a tremendous advantage over the manual effort required by standard methods. Thus, to validate the effectiveness and success of the proposed streamlined strategy for chemical shift assignment analysis, the majority of the results presented below were obtained using spin-system lists generated through the application of this peak-grouping script to either SMILE or Sparky peak lists, requiring very little to no manual analysis of the resulting spin-system lists. We employ this method both for the test case of GB1 as well as a much more challenging case of AL-09 V_L fibrils, and describe the specific types of spectra used and how the peak lists for each were produced. The Python code for the peak-grouping script may be obtained upon request (which can be applied to NCACX, NCOCX, and CANcoCX 3D spectra as well as CANCOCX and CONCACO 4D spectra), and a detailed overview of the steps involved in the peak grouping function can be found in the Supplemental Materials 5.7.1.2.

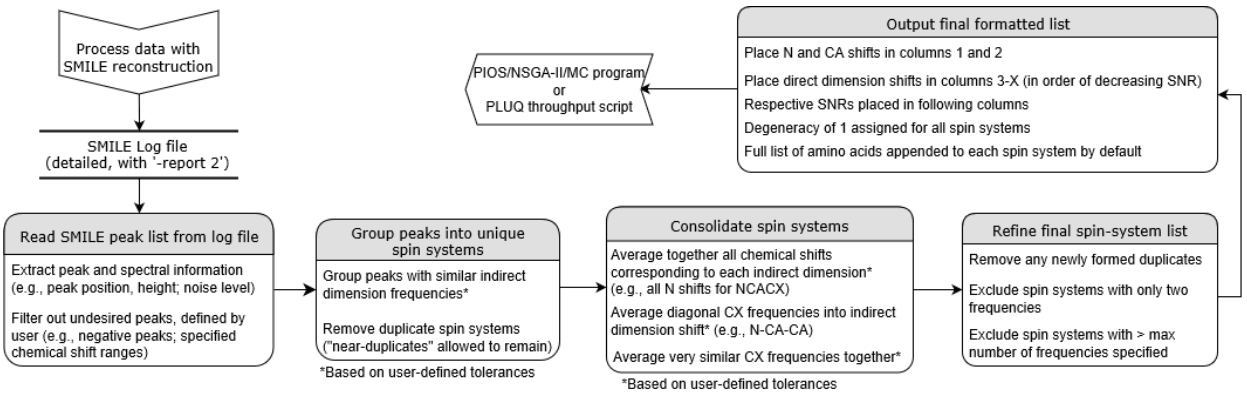

Figure 5.5. Flowchart for the peak grouping script used to generate spin-system lists from SMILE log files for input into the PIOS NSGAII/MC program. Asterisks indicate that the step can be controlled by user-defined input settings.

Last, we also present a script to facilitate the process of determining amino acid guesses for each spin system by throughput into the PLUQ program developed by Fritzsching et al.[28]. Accordingly, this functions to simply take the output of the peak-grouping script above, and depending on the type of spectrum for which the spin-system list represents, extracts specific columns of frequencies and submits them to PLUQ. The output of the amino acid queries is then read back in and appended to the end of each spin-system list, while retaining the proper formatting of the list for input into the PIOS-weighted NSGA-II/MC program. One additional feature that we have included in this step as a way of further assessing the robustness of the presented algorithm is the subsequent expansion of each set of amino acid guesses for each spin-system such that groups of amino acids which exhibit similar chemical shifts will be added onto the given set of guesses (e.g., if a Glu guess is present, guesses for Asp, Gln, and Asn will be included as well if not already present). This serves to increase the likelihood that the correct amino acid guess is included for each spin system, as the exclusion of the correct guess will prevent the spin system from being tested for that particular assignment. Overall, our intent in providing these alternative strategies to the assisted assignment process is to help streamline the tasks that otherwise become more cumbersome to perform, especially when the user is required to also manage and organize the corresponding intensity

information associated with each peak, retain a specific order of frequencies in the final input list for connection purposes, as well as repeat this process for a large number of different input files obtained from spectra processed using a variety of apodization parameters. This PLUQ throughput script is written in Python and may be obtained upon request; a description of the functions performed by the script can be found in the Supplemental Materials 5.7.1.3.

5.5. RESULTS AND DISCUSSION

5.5.1. Validating and assessing the performance of the PIOS-weighted NSGA-II/MC algorithm relative to other methods.

Prior to performing our assignment calculations using the PIOS-weighted NSGA-II/MC program, we carried out a set of control experiments using the input spin-system lists and NSGA-II/MC program published previously[21]. In particular, we applied the original NSGA-II/MC program to the three cases of incomplete GB1 NCACX and NCOCX data sets in which spin systems from each list were deleted in a consecutive, random, or random and independent manner. These lists were prepared artificially with the intent of testing the program's ability to correctly assign the provided regions of the protein sequence despite the possible absence of cross peaks for highly dynamic portions of the protein structure, as can often be the case in other protein systems. In all cases, the NSGA-II/MC program consistently reproduced the successful results previously reported (data not shown). Subsequently, we performed all three tests using the PIOS-weighted NSGA-II/MC program with same input spin-system lists but setting the global chemical shift uncertainties to the average of the corresponding frequency from the provided lists and setting the SNR for all peaks to a moderate value of 10. Our results revealed similarly accurate chemical shift assignments for the regions of the protein sequence which were retained in the input spin-system lists (Figure 5.10), demonstrating that the replacement of the "bad" objective with the PIOS objective does not negatively impact the quality of results obtained by previous versions of the presented assignment program.

While this test provides a strong validation of the program's ability to assign only the residues which exhibit cross-peak intensity, one limitation that it faces is the use of spin-system lists containing fewer than the number of residues present in the protein sequence. As such, an important alternative test of this capacity is to provide the program with spin-system list containing more than the number of residues in the protein sequence, in order to determine if the program will still be able to correctly assign the correct residues despite the additional presence of possible noise, artifacts, or spin-systems arising from the existence of multiple conformations for a single residue. One indirect way of testing this is through the provision of GB1 spin-system lists containing more than the number of expected spin systems (i.e., 56 for GB1) without prior knowledge of whether the all of the truly representative spin systems of the protein are included, as would occur in the case of generating lists by automated means rather than manual analysis.

To pursue this investigation, we proceeded to compare the ability of the PIOS-weighted NSGA-II/MC algorithm relative to the previous NSGA-II/MC program to assign the chemical shifts of GB1 from spin-system lists generated by the streamlined workflow described in the previous section (Figures 3 and 4). Specifically, the peak lists obtained from the SMILE reconstruction of NCACX 50 ms (22% NUS) and NCOCX 50 ms (17% NUS) 3D data as well as CANCOCX 50ms (4% NUS) 4D data were provided as input to the peak grouping script to generate a list of spin systems for which amino acid typing could be performed using the PLUQ throughput script and manual inspection of the results, providing a final spin-system list for each spectrum for input to the PIOS-weighted NSGA-II/MC program. The final set of spin-system lists for all three spectra are included in the Supplemental Materials 5.7.3, which also include the SNR information corresponding to all peaks within each spin system. One notable detail that can be observed in the GB1 input lists is that several of the spin systems contain more than the number of expected frequencies for a particular residue type, particularly among several glycine spin systems (recognized by the unique CA chemical shift frequency). These additional frequencies are the apparent consequence of one or more inter-residue cross peaks occurring within a given spin system and/or may represent the presence of spin-system degeneracy between two separate residues containing similar

frequencies in the ω1 and ω2 dimension for 3D spectra. As we are interested in establishing a more robust program that can successfully handle spin-system lists in which the identification of particular atom types for each frequency is not required (i.e., requiring the user to know in advance what type of side-chain atom a particular frequency corresponds to and thus place in a specific column order for connecting to the other signal lists), the order of ω3 frequencies for each input list that we used for our tests is simply based on decreasing SNR value and does not necessarily reflect the atom type to which each frequency corresponds. Nonetheless, the value of including the SNR value for each frequency becomes most apparent in cases of such spin-system degeneracy or inter-residue correlations, as the intensity of these indirectly associated cross-peak correlations is notably lower than those corresponding directly to the spin-system or residue of interest. Accordingly, the PIOS-weighted NSGA-II/MC program is far less susceptible of being negatively impacted by the presence of such additional frequency information than the previous program which does not account for the relative extent of peak intensity and overlap.

Our initial performance tests were carried out in ten replicates (each with a different input seed number) using only the MC component of the algorithm (5×10^5 attempts) over 30 total steps with an initial group size of 50. Minimum and maximum weights for good, edge, used, and PIOS objectives were set to 0–5, 0–1, 0–1, and 0–2, respectively, while no weight was placed on bad connections in order to avoid penalizing the overall score for each assignment due to the presence of non-representative chemical shifts in several of the spin systems as described above. Additionally, as all possible combinations of connections between all ω3 frequencies of one spin-system list with those of the other spin-system were specified as test connections in the connection file, the inclusion of weight on bad connections can interfere with the program's ability to identify the optimal solution. The input control file containing all details for the runs as well as the expanded connection file are included in the Supplemental Materials 5.7.4 and 5.7.5. Overall, as shown in Figure 5.6A, the PIOS-weighted NSGA-II/MC program performed consistently well over the course of all ten runs, providing accurate chemical shift assignments for the entire GB1 sequence, including the ^{13}C chemical shifts for residue M1 (the N-terminal ^{15}N chemical shift

is folded in the ω1 dimension of the NCACX experiment and thus exhibits the incorrect frequency in the final output, as it was not unfolded during the generation of spin-system lists). To assess how these results compare to those obtained from the exclusion of the PIOS objective, we performed ten independent runs using the original NSGA-II/MC program under the same calculation settings[21], and removing the additional SNR information from the spin-system lists. The percentage of times each residue was assigned correctly across all ten runs for the NSGA-II/MC program is plotted in Figure 5.6B. Notably, the PIOS-weighted program was found to perform statistically better than the original NSGA-II/MC program when provided with the same input spin-system and connection files.

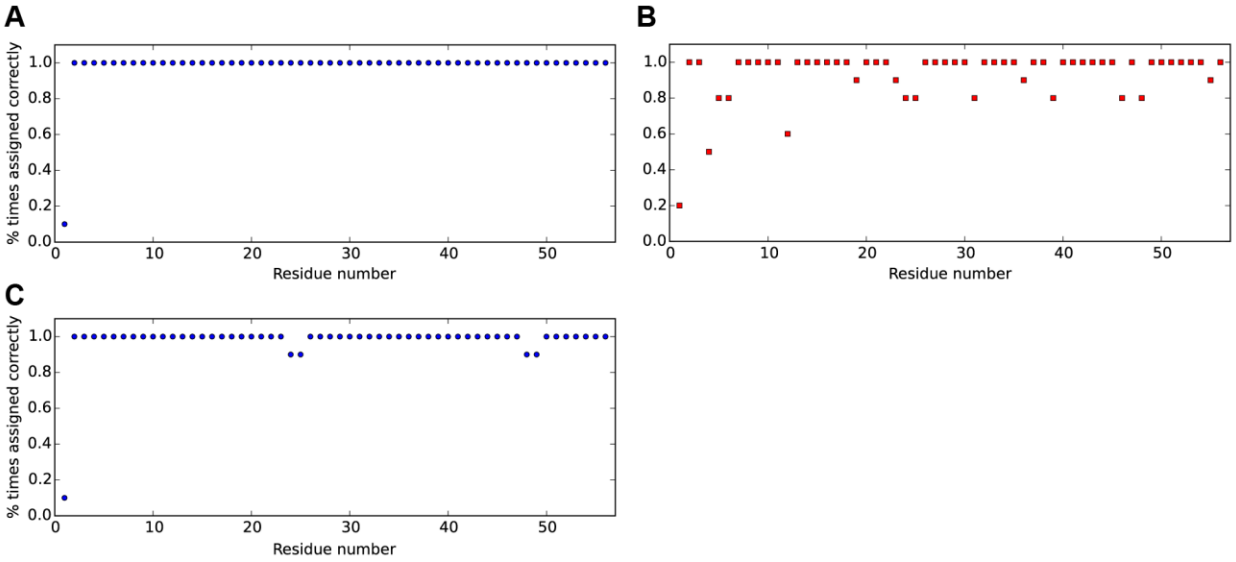

Figure 5.6. Average assignment accuracy as a function of residue number over ten independent simulations on GB1 spin-system lists generated from the grouped output of SMILE reconstructed NCACX 50 ms and NCOCX 50 ms 3D NUS spectra and CANCOCX 50ms 4D NUS data, using only MC steps with the (A) the PIOS/NSGA-II/MC or (B) NSGA-II/MC assignment program. (C) Results over ten runs from the PIOS/NSGA-II/MC program using both MC and NSGA-II attempts. Spin-system lists and the corresponding connection file used for the calculations are included in the Supplemental Materials 5.7.3.

This result provides a clear demonstration of one benefit from including the additional PIOS weight into the computationally-aided assignment process, as the deficiencies of the original program may be attributed to its dependence on high quality input spin-system lists which do not contain any overlapping or non-representative frequencies within a single spin-system. Additionally, the general dependence of the previous programs on being provided input lists in which the atom types are already known and placed in a specific column order accordingly prevents them from being able to handle more realistic data sets in which the atom types are not known. This restriction arises from the program placing equal weight on all frequencies within a given spin system regardless of the identity of the corresponding atom type or the residue to which it belongs (and even for noise peaks), thus becoming highly susceptible to assignment errors caused such as a false positive. This limitation is addressed through the PIOS objective by providing a way to scale the quality of a connection based on the SNR corresponding to each frequency and working under the assumption that overlapping inter-residue peaks within a spin system will exhibit a lower intensity than those that directly belong to the spin system.

Another clear benefit offered through the addition of the PIOS objective is related to the "all-good" or "all-bad" decision making in the original program. Similar to the case described above, if an additional resonance is present in a particular spin system (e.g., the inter-residue correlation of the i+1 $^{13}C\alpha$ chemical shift), the NSGA-II/MC program will identify this as a bad connection and thus assign a lower Pareto order to the given assignment. However, in the case of the PIOS-weighted NSGA-II/MC program, such a connection will either simply be ignored (if no other frequency in the connected spin system overlaps with the inter-residue correlation frequency) or given a relatively lower PIOS value as opposed declaring the connection as being "entirely good" (under the assumption that the inter-residue peak will exhibit a weaker signal intensity relative to an intra-residue correlation). As such, the heavy dependence on an either all-or-nothing objective is alleviated by the inclusion of the PIOS objective, thus preventing the possible elimination of a large number possibly valid solutions resulting from the identification of a single bad connection, as can occur under the design of the original programs.

We also performed a similar set of simulations with the addition of the NSGA-II component as well (5 x 10^3 NSGA-II attempts; 5 x 10^5 MC attempts; 15 SA iterations; max w_1, w_2, w_3, w_4, and w_5 of 2.5, 0.0, 0.5, 0.5, and 1.25, respectively; see Supplemental Materials 5.7.4 for control file). The results from the simulations performed with the PIOS-weighted NSGA-II/MC program demonstrated nearly perfect results for all runs with the exception of four residues which were assigned correctly for 9 of the 10 simulations (Figure 5.6C). We also performed these simulations using the original NSGA-II/MC program for comparison to the presented program; however, due to the inability to turn off the tabulation of bad connections in the previous program, we recognized that this caused severe complications in the sorting algorithm in the NSGA-II attempts and thus led to significant deficiencies in the results (even with the w_3 weight for bad connections set to zero), and thus we have excluded this comparison.

Another noteworthy observation is that in nearly all runs of GB1, the solutions with the highest PIO score represented the one with the most accurate number assignments, even between other Pareto-order-1 solutions that with the same number of "good" connections. Thus, the addition of the PIOS objective provides a valuable complement to the "good" objective and others, as it can offer and additional standard by which the quality of a solution can be assessed. We have found this trend to be particularly useful for tracking the progress and success of an assignment calculation, as it can provide indirect insight on whether—and at what step—the sampling of the program has fallen into a local minimum. Importantly, however, in light of this observation we would also like to point out that due to the way the PIOS objective must be calculated in the program, the overall quantity can become exceptionally large relative to the other objectives, often adding up into the range of tens of thousands (though which is divided by 100 at the scoring step), which can very easily drive the calculation in to a local minimum if the corresponding weighting factor is not set properly. Provided this, we have found the use of 0 as an initial weighting factor and maximum values between 20 to 500 times lower than the number of total steps (e.g., for 50 steps, we would use a $w_5$max of 0.5 to 1.25), but should be adjusted and tested to identify a max value that works well for the particular case.

5.5.2. Evaluating the robustness of the PIOS-weighted NSGA-II/MC program with unanalyzed and computationally generated spin-system lists.

We next aimed to investigate the robustness of the PIOS-weighted NSGA-II/MC algorithm by providing it with more challenging spin-system lists. While the lists provided for the earlier tests were observed to contain several inter-residue correlations for a few spin systems, those lists were generated using a high sigma level in the processing of the data with SMILE and as such does not contain many of the relatively weaker peaks that are still present in the data. As such, we reprocessed the NCACX 50 ms and NCOCX 50 ms 3D NUS data sets using a lower sigma threshold of 10, and provided the output log files and input to the peak-grouping script work flow. This provided us with two new NCACX and NCOCX spin-system lists that contained 95 spin systems and 73 spin systems, respectively. Further, in order to avoid the possibility of excluding the correct amino acid guesses for the relevant spin systems of GB1, we specified in the PLUQ throughput script to include all amino acids of the following groups if one is present in the set of guesses for a given spin system: (1) I, V, and T; (2) D, N, E, and Q; (3) K, L, and I; and (4) F, W and Y.

Subsequently, we performed a series of 5 independent runs using these lists in combination with the CANCOCX list used for the previous simulations, with different initial control files settings. The first set of simulations were performed using only MC attempts and over 30 full program (SA) iterations. As shown in Figure 5.7A, nearly all of the residues of GB1 were assigned correctly in the majority of the simulations. We next performed a similar set of 5 independent runs using both MC and NSGA-II attempts, but with only 15 total SA iterations. Interestingly, the results from these simulations demonstrated significantly lower agreement with the correct GB1 chemical shift assignments (Figure 5.7B). Yet, as we describe in the results on AL-09 V_L fibrils in one of the following sections, the NSGA-II component does not necessarily consistently lead to lower quality results and may only have been a consequence of the use of fewer simulated annealing steps here. Alternatively, it may be possible that the NSGA-II algorithm performs less optimally when provided such complicated data sets (i.e., those

containing many non-representative frequencies within a given spin-system). Nonetheless, we have found both the MC/SA as well as the NSGA-II methods to offer their own benefit to the assignment results in a variety of other cases, and recognize them each as valid algorithms to the assignment process.

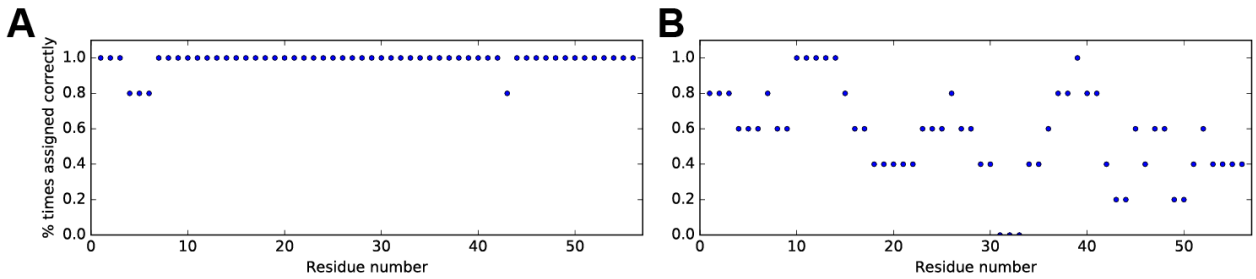

Figure 5.7. Accuracy and precision of PIOS-weighted NSGA-II/MC assignments of GB1 as a function of residue number over five independent simulations using NCACX, NCOCX, and CANCOCX spin-system lists in which many inter-residue correlations are present. (A) Assignment simulations run using only MC attempts (5 x 10^5) over 30 steps with a group size of 50. (B) Simulations run using both MC attempts (5 x 10^5) and NSGA-II attempts (5 x 10^3) over 15 steps. NCACX 50 ms and NCOCX 50 ms NUS 3D lists were generated directly from SMILE reconstructed peak lists (processed using a relatively lower nSigma value of 10 and time domain limits of 0.1 to 1.0) and PLUQ throughput, followed by automated addition of similar amino acid types.

One particularly interesting example of the success of robustness for the PIOS-weighted algorithm can be seen in the assignment of glycine 9, in which multiple inter-residue correlations are observed in the single spin system. As illustrated in Figure 5.3, the $^{13}C\alpha$ resonances for both the i-1 and i+1 residues are present within the spin system. Despite this, the assignment program was still able to successfully assign this residue in every simulation with only MC attempts (Figure 5.7A), and for the majority of runs with both MC and NSGA-II attempts (Figure 5.7B). Intriguingly, after reviewing the assignment solution resulting from one of the fully accurate simulations, we observed that the spin

systems of N8 and K10 in the NCACX list likewise contain the $^{13}C\alpha$ frequency corresponding to G9 (44.8 ppm). Although the connection file that we used did not specify the comparison of frequency column #2 ($^{13}C\alpha$ for the NCACX list) with any ω_3 frequencies in the other lists—and thus preventing the connection of this frequency with its presence as an inter-residue correlation in another spin system—this raises the consideration of whether doing so might also be of value, as the observation of such neighboring correlations when analyzing data manually is typically used as a valuable way to confirm the accuracy of an assignment. Certainly, an investigation on the added value of including potential inter-residue correlations into the connection list is an interest we intend to pursue in future studies. Overall, the results that we have obtained through this method show clear improvement relative to the original NSGA2/MC program, and demonstrate that even if the input spin-system lists contained inter-residue correlations, the PIOS-weighted program is still able to correctly assign the residue for the primary ω_1 and ω_2-containing spin system.

5.5.3. Sequential chemical shift assignments using a single four-dimensional CANCOCX NUS spectrum.

In addition to analyzing the overall robustness of the PIOS-weighted NSGA-II/MC algorithm, we were also interested in assessing the capacity of the program to perform sequential chemical shift assignments with as minimal data as possible. Towards this, we acquired a CANCOCX 50 ms 4% NUS 4D spectrum and a CANCOCA SPC-7 q56 (1.28 ms) 4% NUS 4D spectrum on GB1 and generated a corresponding spin-system list for the data set through the SMILE reconstruction and peak-grouping workflow. At the amino acid guessing step, however, we decided to use the single spin-system list input to generate two separate outputs from the PLUQ processing script—one in which amino acid guesses were made based on the CA and N chemical shifts (ω_1 and ω_2), and another based on the CO and CX chemical shifts (ω_3 and ω_4). Additionally, we prepared a new connection file in which we could relate the two separate spin-system lists by the matching CA and CX frequencies of ω_1 and ω_4. Upon inputting these

files into the PIOS-weighted NSGA-II/MC assignment program and analyzing the results, we found that all but two residues of the protein sequence were assigned correctly (Figure 5.8). Surprisingly, the only residues that were misassigned (N8 and A20) are not notably degenerate spin-systems out of the list of GB1 assignments, and thus the program's inability to correctly assign these residues in particular was not expected. Nonetheless, the ability to assign nearly all other residues with only a single, unanalyzed NUS 4D spectrum raises tremendous interest in the possibility of applying this to more difficult systems and further improving the standard for assigning proteins in a more efficient and robust manner.

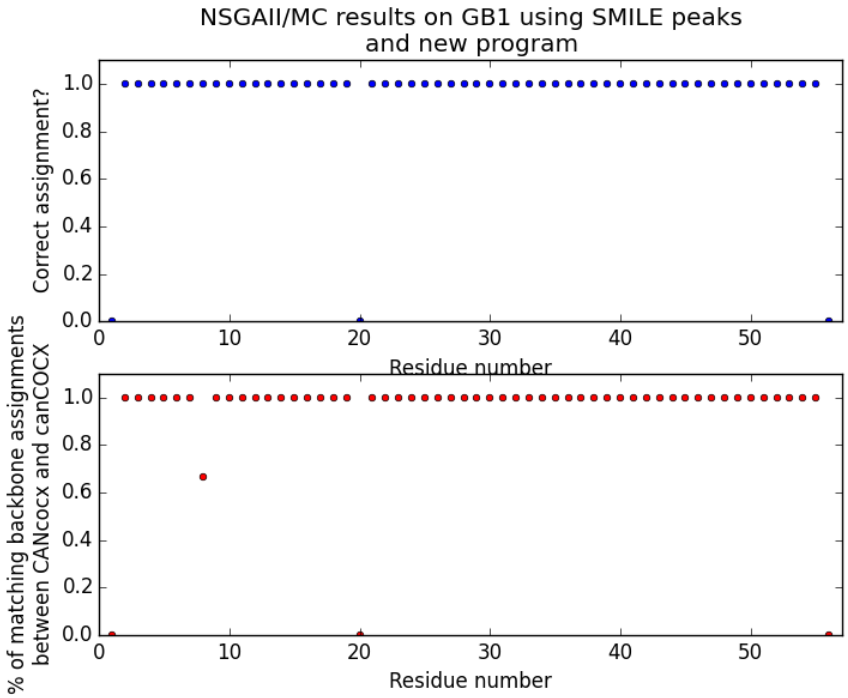

Figure 5.8. PIOS/NSGA-II/MC assignment results for GB1 using a single spin-system list generated from the combined grouping of peaks reconstructed by SMILE for a CANCOCX 50 ms NUS 4D spectrum and CANCOCA SPC-7 q56 NUS 4D spectrum.

5.5.4. Augmenting chemical shift assignments of AL-09 V_L fibrils with PIOS-weighted NSGA-II/MC.

After validating the success and robustness of the PIOS-weighted NSGA-II/MC assignment program using the test case of minimally edited SSNMR data of GB1, we proceeded to extend our approach to more challenging protein systems which exhibit a higher extent of spin-system degeneracy, peak overlap, as well as spectral congestion for proteins of larger size. One representative example of this is exhibited by the case of AL-09 V_L fibrils, for which we have recently reported the backbone and side-chain chemical shifts as well as insight on the secondary structure and dynamics of different regions of the fibril[29,30]. The protein sequence for AL-09 V_L is 110 amino acids, and exhibits a disproportionately high number of serine and threonine residues (14 Ser and 12 Thr) relative to other amino acid types, and the presence of which we have observed to contribute to partial peak overlap due to similarities in backbone and side-chain chemical shifts. Several other notable features of the AL-09 V_L protein sequence that present substantial assignment challenges are the occurrence of repetitive amino acid motifs (e.g., $_{63}$SGSGSGT$_{69}$), as well as the presence of 8 proline residues throughout the sequence. While proline spin system generally consists of a set of rather unique chemical shift frequencies relative to other amino acids (e.g., a ^{15}N shift typically ≥ 130 ppm and ^{13}Cα shift ~60 ppm), it poses a notable challenging case in ^{15}N-^{13}C correlation SSNMR spectra, as the lack of an amide proton results in a poor overall signal for the proline spin system in (^{1}H)-^{15}N-^{13}C-^{13}C type experiments that rely upon ^{1}H-^{15}N polarization transfer for the initial signal.

The spin-system lists used for performing the automated assignment program on AL09 were prepared in the following manners: (1) an NCACX spin-system list was obtained by combining the SMILE output peak lists from a NCACBCO SPC-10 q64 (25% NUS) 3D spectrum reconstructed using nSigma values of 6 and 8; (2) NCOCX spin-system list assembled from the combination of several manually analyzed NCOCX and NCOCA 3D spectra; and (3) a CANCOCX spin-system list prepared from the combination of SMILE reconstructed peak lists for a CANCOCX 50ms 4D (2% NUS) spectrum

and CANCOCA SPC-8 q64 4D (4% NUS) spectrum, using an nSigma value of 4. All three of these grouped spin-system lists were then provided as input to the PLUQ throughput script to obtain an initial set of amino acid type guesses for each spin system.

Using these spin-system lists as input the PIOS-weighted NSGA-II/MC program, we performed 50 independent simulations using either only MC attempts (50,000) or both MC and NSGA-II attempts (50,000 and 5,000, respectively), and all runs were performed using 30 total steps, a group size of 50, gene pool size of 40, minimum SA weights of 0, and maximum weights of 5, 0, 1, 1, 1.25 for good, bad, edge, used, and PIOS, respectively. The results from the simulations are plotted in Figure 5.9 for the MC-only (blue) as well as the combined MC and NSGA-II (red) runs as a histogram of the number of times a residue was assigned correctly out of all 50 runs (i.e., matching with manual assignments) as a function of residue number. Interestingly, both initial calculation setups performed with very similar success on average, which was unexpected in light of the previous test of GB1 earlier, in which the combined MC and NSGA-II simulations produced much lower quality results than the MC only simulations. As such, it is possible that the reason for lower success for GB1 may have been due to the use of an insufficient number of total SA steps, as for the case here the same number (30 steps) was used in both cases. Certainly, this is an additional test we will perform briefly. Beyond the similarity of MC and MC with NSGA-II outputs, one noteworthy consistency between the two is that in both simulation setups, the N- and C-terminal residues are observed to be assigned correctly more frequently than the other residues, which further augments our previous chemical shift assignments as well as our observations of highly ordered and resolved spin systems for residues within these regions[29,30].

One interesting observation in the results shown in Figure 5.9 is that several of the residues were not assigned correctly in any of the total 100 simulations, specifically: Q3–M4, P8, C23, V58–P59, R61–F62, and F98. While at initial glance this may appear to be signify a deficiency in the PIOS-weighted NSGA-II/MC program, re-inspection of the generated input lists revealed that most of these residues were reconstructed by SMILE and grouped accordingly into the input spin-system lists; however, the limiting

factor in this situation was most often the absence of the accurate amino acid guess predicted for the given set of chemical shifts by PLUQ. Nonetheless, one way of addressing this shortcoming would be to perform multiple variations of the simulations using several different versions of grouped spin-system lists (e.g., using different SMILE parameters, using different peak-grouping tolerances, as well as different types of spectra), which is a general recommendation we make when performing computationally-assisted assignments. Indeed, in using a variety of spectra types and processing parameters—in combination with the use of the streamlined approach we describe and apply here—one may be much more easily able to recognize the benefit of inputting ample (though perhaps not "perfect") data into the assignment procedure as an alternative to a single manually-refined data set.

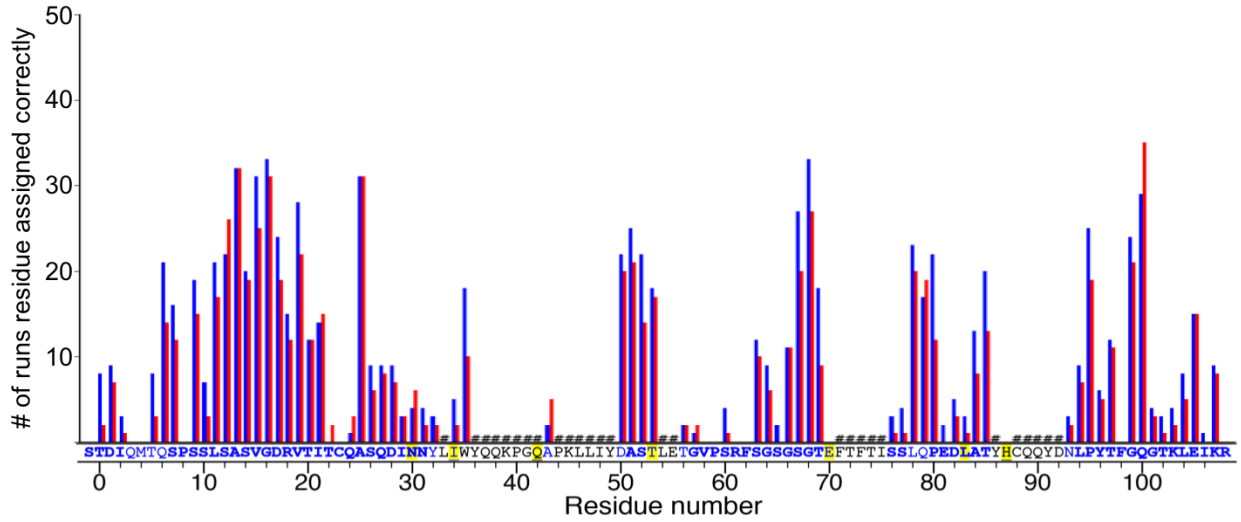

Figure 5.9. Comparison of the consistency of PIOS/NSGA-II/MC results between 50 different simulations with manual assignments of AL-09 V_L fibrils, using either only MC attempts (5×10^5; blue) or using both MC and NSGA-II attempts (5×10^5 and 5×10^3; red). NCACX and CANCOX spin-system lists generated from grouping of the reconstructed peak list output from SMILE as well as a NCOCX spin-system list generated from manual assignments in Sparky. Reconstructed NCACX and CANCOCX spectra include a 25% NUS NCA(CB)CO 3D spectrum with SPC-10 q60 mixing, a 4% NUS CANCOCA 4D spectrum with SPC-8 q64 mixing, and a 4% CANCOCX 50ms 4D with J-decoupling soft pulses in both indirect ^{13}C-dimensions. Hash-mark symbols (#) indicate that no manual assignment has been made and thus was not included in the plot.

5.6. CONCLUSIONS

We present and demonstrate the improved performance of a robust assignment program that uses peak intensity and frequency overlap information for scoring the quality of set of given assignments, which we call the PIOS objective. In addition to showing that this method functions with consistently high accuracy for assigning the chemical shifts for the test case of GB1, we demonstrate that the PIOS-weighted NSGA-II/MC program can produce similarly successful results for input data sets containing a high number of inter-residue correlations, and does not rely upon any specific ordering of chemical shifts in the input files. To this end, we apply this algorithm to a streamlined workflow in which no manual analysis of the data is required. In addition to showing success with GB1, we also demonstrate appreciable assignment accuracy for more challenging cases of AL-09 V_L fibrils which exhibit a higher degree of spin-system degeneracy and peak overlap. Overall, the presented PIOS-weighted NSGA-II/MC algorithm is demonstrated to offer substantial improvement relative to the previously reported algorithms which do not take into account the peak intensity information as a way to guide the assignment process.

5.7. SUPPLEMENTAL MATERIALS

5.7.1. Supplemental information

5.7.1.1. Notes on the modifications implemented in PIOS-weighted NSGA-II/MC program source code

The source code for the PIOS-weighted NSGAII/MC program is written in Fortran 95, and is an extension of the NSGA-II/MC code written by Yang et al.[21] as well as the MCASSIGN2 code from Hu et al.[20], respectively. The corresponding Fortran 95 files for generating the PIOS-weighted NSGA-II/MC executable as well as a template Makefile are may be obtained upon request. In the implementation of the PIOS objective into the program, several modifications to the previous code were necessary in order to accommodate the additional quality score calculation in addition to several other minor changes. Of these minor modifications includes our decision to change the declaration of connections that fall outside the limits of the protein sequence (e.g., an NCOCX spin system assigned to the C-terminal residue) from bad

connections to be edge assignments instead, since we believe this better represents the meaning of the connection and that an impossible assignment of this type should not be penalized with a higher weight (as w2 is typically set to be higher than w3).

Other small changes were generally gated around addressing Fortran compilation issues, through which we have now made the program fully compatible with the freely-available PGI community edition compiler. This compiler is available for download on the PGI website for OS X, Unix, and Windows operating systems, so long as the system requirements for installing the compiler are met. We have confirmed that the PIOS-weighted NSGAII/MC program compiles and runs successfully on Mac OS X 10.10, Ubuntu 16.04, and CentOS Linux 7.3 working environments. The executable used here was generated using an optimization level of 2, with the added debugging option flag, '-g'.

5.7.1.2. Description of the functionality and logic of the peak-filtering and grouping script

The peak filtering and grouping script is written in Python and may be provided upon request. It can be run simply through the command line and requires only the installation of the NumPy module. The script functions by first reading in a Sparky peak list file or SMILE log file generated from the processing of a NCACX, NCOCX, or CANcoCX 3D spectrum or CANCOCX 4D spectrum using SMILE reconstruction on either NUS or US data (with the '-report 2' flag, for a detailed log file), and extracting out all of the peak information as well as the thermal noise level. Additionally, depending on the settings specified by the user, the script will serve as a filtering function at this step by either retaining or excluding negative cross-peak signals, as well as peaks that fall outside certain indirect dimension chemical shift ranges (e.g., we retained peaks within ^{15}N ranges of 95–150 ppm, ^{13}CA ranges of 38–70 ppm, and ^{13}C' ranges of 165–185 ppm). We chose to exclude negative peaks from data sets acquired with DARR mixing as they did not represent peaks of interest, but they were retained for experiments performed using SPC-N dipolar recoupling schemes, which result in negative cross-peak intensities for directly coupled ^{13}C-^{13}C correlations.

Next, peaks with similar chemical shift frequencies in the indirect dimensions (i.e., w1 and w2 for 3Ds; or w1-w3 for 4Ds) are grouped together based on user-defined tolerances (in ppm), forming an initial set of spin systems. Typical tolerances used for the spin system lists generated here were 0.25 ppm for ^{15}N dimensions and 0.15 ppm for ^{13}C dimensions, which were chosen to cover the full range of chemical shifts within the standard deviations of the average shift value often observed for the corresponding nucleus type in SSNMR spectra (i.e., 0.5 ppm for ^{15}N; 0.3 ppm for ^{13}C). As this grouping process can lead to the generation of redundant spin systems (e.g., peaks A and B will be grouped, as will peaks B and A separately, in the format by which it is currently implemented), these duplicates are removed; yet, it is also worth noting that while peaks A and B are grouped based on the indirect dimension frequencies of A, there may be an event where peak B is later grouped with both A as well as additional peaks C, D, etc. In such an event, both spin systems are retained in the final output signal list, even though they may differ by only one frequency, or by the averaged shift of one of the frequencies. Although this may be undesirable in some situations—becoming most problematic at the assisted-assignment stage due to increasingly longer calculations—we believe this presents as a highly suitable solution to the question of possible spin-system degeneracy. In other words, rather than requiring the user to assess whether a particular spin system should be given a degeneracy value greater than 1, the generation of multiple, similar—yet slightly different—sets of spin systems can serve as an alternative approach to the decision-making process, and as such the script by default assigns a degeneracy value of 1 to every generated spin system in the final output list. This of course can be modified later by the user, if desired. Another added and interesting benefit to the possibility of redundant spin-system grouping is that the PIOS-weighted algorithm will, in the end, identify the spin system which agrees best with the other spin-system lists based on the slight differences in chemical shift frequencies (as this is accounted for in the overlap score) as well as SNR values of the corresponding peaks selected to be grouped in each spin system. As such, for the assignment calculations performed in our work here, all of the spin systems generated by the grouping script were retained to provide the algorithm with the complete set of data to use for assignments, rather than unintentionally excluding solutions that may be more optimal.

Following the initial filtering and grouping steps, all of the indirect dimension frequencies are separately averaged within each spin-system group, providing a single average chemical shift value for the ω1 and ω2 frequencies of the spin system (and ω3 for 4Ds). Further, as peaks occurring at the diagonal in a 3D or 4D spectrum will exhibit nearly the same chemical shift value in both the direct dimension and one of the indirect dimensions (e.g., an N-CA-CA correlation), the script will average the direct dimension chemical shift into the corresponding indirect dimension based on a separate user-defined tolerance range (we used 0.15 ppm for ^{13}C shifts, which represent the only nucleus for which a diagonal was present in the data analyzed here). Subsequently, a similar averaging process is performed for all direct dimension chemical shifts within another user-defined ppm range, as we observed that in some instances (based on the settings used in the data processing step), SMILE can overrepresent certain peaks as multiple signals and thus lead to the presence of many very similar chemical shifts within a single spin system. During this process, a decision is made on how to handle the SNR values corresponding to the two peaks for which the direct dimension chemical shift will be averaged, in which the script will average the two SNR values if the value of the weaker peak is within 50% of that of the stronger peak; otherwise, the stronger SNR is retained to correspond to the averaged chemical shift. Related to this, as the final output list must contain a SNR value corresponding to each chemical shift frequency in the spin system in order to be read in properly by the PIOS-weighted NSGAII/MC program, there must be an SNR assigned to the indirect dimension frequencies as well. Thus, we decided the most appropriate way to address this is to use the max SNR value from within each spin system to represent the SNR of the indirect dimensions.

Last, the script sorts the grouped spin systems based on the ^{15}N chemical shift frequency, in addition to sorting the direct dimension chemical shifts within each spin system based on the strength of each frequency's corresponding SNR, before outputting the list as a new text file. This ordering of direct dimension shifts is generally trivial, but is done so that the subsequent PLUQ throughput script will select the stronger frequencies to use as input for amino acid querying, as only the first five frequencies are used for each query. The final output signal list contains a header line at the top specifying the number of

grouped spin systems present in the file as well as the number of frequencies included within each spin system (we used maximum sizes of 9 frequencies per spin system; however, this can be specified by the user when running the script), followed on the next line by the estimated chemical shift uncertainties (in ppm) for each corresponding frequency (calculated as 2*tolerance of each nucleus type as specified by the user in the grouping script). The remaining rows of the file represent the grouped spin systems, and are printed out in the following column-order (for n frequencies):

^{15}N (ppm) | ^{13}CA (ppm) | ^{13}C'/^{13}CX$_{1–n}$ | SNR ^{15}N | SNR ^{13}CA | SNR ^{13}C/CX$_{1–n}$ | Degeneracy | AAs

We note here that the output spin-system list from this script will be formatted properly for the PIOS NSGAII/MC program, but by default will contain the full list of possible AAs for each spin system (i.e., all 20 amino acid letter codes will be filled into the "AAs" column field). This is done primarily to serve as a placeholder for until the frequencies from each spin system in the list can be run through PLUQ. Because the process of filtering and grouping the peak list is a much faster computational task than the I/O involved in querying a long list of spin systems, we decided to make the PLUQ throughput script as a separate step than incorporating it into the grouping job. This allows the user to assess and review the resulting spin system list before running it through PLUQ, and make any changes to the tolerances or indirect dimension window limits (or to the final list itself) that they wish to make without having to wait several seconds to a few minutes (depending on the length of the list) between each change.

5.7.1.3. Description of the functionality of the PLUQ throughput script

Once the user is satisfied with the resulting list of spin systems for the particular spectrum of interest, this list may be used as input for a separate PLUQ[28] throughput script, to obtain the top amino acid guesses for each spin system. This script works by using user-specified chemical shift columns from each spin-system list as frequencies to input into the PLUQ program, and essentially concatenating the

input command lines to PLUQ, which are then submitted to the terminal and the terminal output is saved to a separate file. The script then reads in the results from PLUQ, converts all amino acids to single letter type, and then appends the guesses to the end of each spin system prior to final output to the proper format for running in the PIOS NSGAII/MC program. This script only requires that the user has PLUQ installed on their system, and specifies the path to the PLUQ directory within in the Python throughput script. This PLUQ throughput script, as well as the peak filtering and grouping script, may be obtained upon request.

In this manner, by streamlining the steps involved in generating input files for the PIOS NSGAII/MC program, our intent is to give the user the ability to explore another frontier in the realm of possible assignment solutions—that of processing-dependent variations in reconstructed data. As such, we suggest that the user process the data in multiple different ways, as well as filtering and grouping the script with several different tolerance settings, so that they have multiple slightly different versions of spin-system lists from the same data set to use in the PIOS NSGA-II/MC program. As the top scoring solutions will receive Pareto-order 1 ranking, one may compare the results from each of these and assess which chemical shifts are identified most frequently for each residue.

5.7.1.4. Description of the functionality of the PIOS-weighted NSGA-II/MC post-processing script

To facilitate the analysis of results from the PIOS-weighted NSGA-II/MC program, we also provide a Python script for averaging the chemical shifts for the top scoring solutions, and is essentially a direct re-write of the MATLAB scripts provided by Yang et al.[21]. This can be provided upon request. One caveat to be aware of when using this script, however, is that if a connection file involving multiple CX to CX connections is used, the results from this script may likely represent averages of non-similar chemical shifts. For example, the CX1 shift of the NCACX list may be averaged with the CX2 of the NCOCX list, if the connection file specified the possible connection between the CX1 and CX2 between the two lists, as this script uses the information in the connection file to determine which chemical shifts to average among the tops scoring solutions. Nonetheless, so long as no degenerate connections are specified for the

^{15}N and ^{13}CA chemical shifts (as we have avoided in our connection lists), the average chemical shift values for these resonances should be represented properly.

5.7.2. Supplemental figures

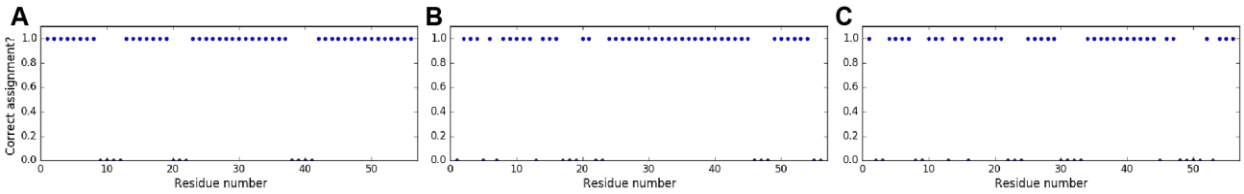

Figure 5.10. Validation of the ability of the PIOS-weighted NSGA-II/MC program to provide consistently accurate chemical shift assignments as the NSGA-II/MC program in the case of GB1 for which 20% of the NCACX and NCOCX spin-system lists provided as input are deleted (A) consecutively, (B) randomly, and (C) randomly and independently. The input files used for these calculations were prepared from the example data provided by Yang et al.[21], in which the only minor changes were made to adapt them to the formatting requirements of the PIOS-weighted NSGA-II/MC program. The SNR for all frequencies was set to 10, and the global chemical shift uncertainties for each frequency column were set to the average of the uncertainties reported for each individual spin system in the original lists.

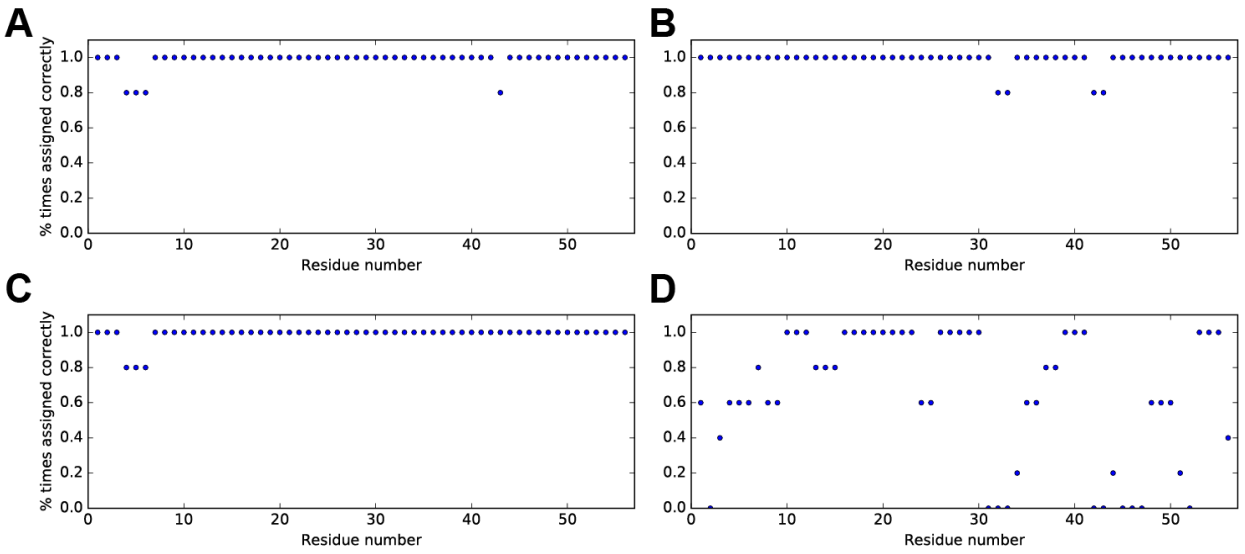

Figure 5.11. Assessing the influence of SNR values and the presence of inter-residue correlations on PIOS-weighted NSGA-II/MC assignment success. Simulations were performed over 5 independent runs using (A) the original NCACX, NCOCX, and CANCOCX spin-system lists, which contain many inter-residue correlations, (B) the same lists with all SNR values set to 30 for all spin-systems, (C) with the SNR values corresponding to the first three frequencies of the *i*th residue (i.e., those with a shift index of 0 in the connection table) set to 10 and all others set to 30, and (D) with the SNR values corresponding to shifts 1–3 in the NCACX and NCOCX set to 5, shift 4 in the NCOCX set to 10, and shifts 1–3 in the CANCOCX set to 5, 5, and 3, respectively, while all others were set to 30.

5.7.3. Supplemental tables: NCACX, NCOCX and CANCOCX spin-system lists for GB1, without extensive inter-residue correlations. Lists were generated from output peak lists from SMILE reconstruction of NCACX 50 ms (22% NUS) and NOCX 50 ms (17% NUS) 3D data as well as CANCOCX 50ms (4% NUS) 4D data with the peak-grouping script.

5.7.3.1. NCACX List

67	9																		
0.5	0.4	0.3	0.3	0.3	0.3	0.3	0.3	0.3											
104.071	69.866	21.607	2000	2000	2000	2000	2000	2000	19	19	19	0	0	0	0	0	0	1TVI
104.081	60.448	21.566	69.898	175.518	2000	2000	2000	2000	65	65	65	57	42	0	0	0	0	1ITV
105.534	45.057	171.074	2000	2000	2000	2000	2000	2000	70	70	55	0	0	0	0	0	0	1G
106.328	62.016	69.42	22.571	173.055	2000	2000	2000	2000	46	46	41	41	35	0	0	0	0	1TVI
108.099	45.253	172.46	2000	2000	2000	2000	2000	2000	92	92	68	0	0	0	0	0	0	1G
108.341	46.931	173.765	61.871	21.886	2000	2000	2000	2000	90	90	64	17	16	0	0	0	0	1G
109.016	61.015	21.085	73.057	173.683	2000	2000	2000	2000	38	38	37	35	29	0	0	0	0	1ITV
109.45	44.777	172.926	2000	2000	2000	2000	2000	2000	68	68	51	0	0	0	0	0	0	1G
112.002	62.581	21.342	71.548	174.146	175.34	2000	2000	2000	52	52	52	48	37	16	0	0	0	1ITV
112.048	60.423	171.953	20.939	71.706	2000	2000	2000	2000	41	41	32	28	17	0	0	0	0	1TVI
114.879	60.25	171.819	70.492	20.06	174.057	2000	2000	2000	85	85	53	52	41	19	0	0	0	1TVI
114.894	53.631	174.036	176.833	40.23	2000	2000	2000	2000	58	58	40	37	36	0	0	0	0	1DNLEQKI
115.359	52.596	179.774	174.958	42.184	2000	2000	2000	2000	41	41	35	30	26	0	0	0	0	1DLNEQKI
116.02	63.685	174.825	31.942	21.201	2000	2000	2000	2000	49	49	36	30	15	0	0	0	0	1PVIKL
116.178	61.446	18.932	70.771	171.054	2000	2000	2000	2000	52	52	45	43	39	0	0	0	0	1ITVS
116.233	59.215	29.047	35.485	181.486	177.605	2000	2000	2000	44	44	37	32	31	30	0	0	0	1HQVKREIDNL
116.239	60.451	174.114	72.489	21.792	2000	2000	2000	2000	44	44	38	34	32	0	0	0	0	1ITV
117.208	60.325	32.716	178.821	26.347	30.012	2000	2000	2000	71	71	53	37	36	26	0	0	0	1MREQKDNLI
117.236	67.464	21.401	175.645	2000	2000	2000	2000	2000	59	59	51	31	0	0	0	0	0	1TVIKL
118.085	57.156	175.997	39.201	179.288	2000	2000	2000	2000	75	75	60	56	49	0	0	0	0	1YDLFNEQKIW
118.141	58.558	21.82	19.725	172.404	32.687	2000	2000	2000	39	39	39	35	30	25	0	0	0	1CITVKL
118.491	57.623	37.602	178.825	2000	2000	2000	2000	2000	17	17	17	16	0	0	0	0	0	1DKFNILEYVTQW
118.515	57.685	37.602	178.825	2000	2000	2000	2000	2000	38	38	17	16	0	0	0	0	0	1DQKFEILNYVTW
118.567	57.791	171.619	42.666	178.825	2000	2000	2000	2000	38	38	19	19	16	0	0	0	0	1DQKFLEYNIW
118.588	57.817	171.619	42.666	2000	2000	2000	2000	2000	38	38	19	19	0	0	0	0	0	1DFILYVNEQKW
118.881	54.104	19.014	179.379	2000	2000	2000	2000	2000	100	100	100	49	0	0	0	0	0	1DKALRENQI
119.003	66.387	21.013	31.841	22.158	178.602	2000	2000	2000	59	59	59	51	47	36	0	0	0	1IVKL
119.226	55.332	31.518	177.621	36.187	45.221	72.582	2000	2000	76	76	47	38	29	14	13	0	0	1MREKDNQLI
119.469	55.616	28.065	175.191	24.333	43.28	2000	2000	2000	33	33	33	31	25	15	0	0	0	1HQKLREDNI
119.494	55.663	24.333	175.191	28.065	43.28	2000	2000	2000	33	33	33	31	25	15	0	0	0	1HKLQRDNEI
119.498	55.675	175.191	24.333	43.28	2000	2000	2000	2000	31	31	31	25	15	0	0	0	0	1DLFNEQKIWY
120.482	54.622	18.218	1.173	2000	2000	2000	2000	2000	123	123	123	45	0	0	0	0	0	1DKALRENQI
120.498	60.188	31.63	179.413	27.331	29.397	2000	2000	2000	47	47	35	30	29	14	0	0	0	1MREQKDNLI
120.75	61.524	178.408	38.678	2000	2000	2000	2000	2000	27	27	25	21	0	0	0	0	0	1IKFELNYVDQW
120.786	54.011	34.236	173.966	2000	2000	2000	2000	2000	59	59	45	41	0	0	0	0	0	1REQKDNLI
120.911	59.332	32.762	178.888	25.689	173.007	29.124	2000	2000	75	75	66	42	41	18	15	0	0	1MREQKDNLI
121.007	56.053	175.962	38.23	177.66	2000	2000	2000	2000	58	58	42	40	33	0	0	0	0	1DKNILEYVTQFW
121.118	58.918	28.842	34.03	177.324	179.731	2000	2000	2000	51	51	51	41	34	26	0	0	0	1REQKDNLI
121.623	61.845	21.904	174.947	31.922	46.92	173.72	2000	2000	96	96	96	40	39	17	14	0	0	1PVKIL
122.146	54.334	30.139	32.362	171.155	2000	2000	2000	2000	44	44	44	43	39	0	0	0	0	1MREQKDNLI
122.456	56.13	18.005	179.388	2000	2000	2000	2000	2000	93	93	93	43	0	0	0	0	0	1DKALRENQI
122.563	54.611	18.256	179.549	174.774	2000	2000	2000	2000	131	131	88	54	18	0	0	0	0	1DKNLAEQI
122.749	54.891	36.155	25.516	28.852	173.067	2000	2000	2000	38	38	38	35	26	15	0	0	0	1EKDNQLI
123.06	53.434	38.784	175.55	26.004	29.697	2000	2000	2000	45	45	35	32	27	14	0	0	0	1DNILVEQK
123.135	57.109	174.638	43.393	2000	2000	2000	2000	2000	29	29	28	18	0	0	0	0	0	1DKFILEYVNQW
123.328	54.674	42.973	179.635	2000	2000	2000	2000	2000	39	39	39	13	0	0	0	0	0	1DLFNEQKIWY
123.368	42.963	179.65	2000	2000	2000	2000	2000	2000	15	15	14	0	0	0	0	0	0	1G
123.695	55.07	17.606	177.082	2000	2000	2000	2000	2000	80	80	71	48	0	0	0	0	0	1KEDLANQI
123.705	55.09	17.477	177.082	2000	2000	2000	2000	2000	80	80	72	48	0	0	0	0	0	1KEDLANQI
123.726	55.111	17.477	2000	2000	2000	2000	2000	2000	80	80	72	0	0	0	0	0	0	1DQKLRENI
123.981	61.389	21.439	72.013	173.857	2000	2000	2000	2000	53	53	49	45	40	0	0	0	0	1ITVKL
124.903	57.571	177.024	33.246	112.339	127.113	2000	2000	2000	33	33	25	21	16	14	0	0	0	1QEVKIDNLW
124.904	50.826	176.303	38.266	2000	2000	2000	2000	2000	46	46	46	39	0	0	0	0	0	1LDINVEQK
125.103	55.905	30.352	174.742	35.204	171.172	2000	2000	2000	66	66	50	45	23	15	0	0	0	1REQKDNLI
125.238	54.452	175.668	30.832	35.931	171.095	2000	2000	2000	67	67	51	43	20	15	0	0	0	1REQKDNLI
125.431	50.867	23.575	177.437	2000	2000	2000	2000	2000	89	89	80	45	0	0	0	0	0	1NDLIAVTEQK
126.067	60.058	17.516	27.345	37.838	175.023	12.717	2000	2000	49	49	49	38	34	30	24	0	0	1IVKL
126.243	51.105	42.71	180.181	176.144	2000	2000	2000	2000	39	39	39	33	26	0	0	0	0	1NDLFEQKIWY
126.89	54.808	174.793	25.097	27.032	2000	2000	2000	2000	28	28	28	27	26	0	0	0	0	1LRKI
126.952	53.041	24.98	27.293	25.777	174.546	42.502	2000	2000	32	32	25	24	23	23	14	0	0	1LKI
127.314	54.606	26.02	43.144	2000	2000	2000	2000	2000	48	48	33	33	0	0	0	0	0	1LRKI
127.359	54.583	43.144	26.021	73.699	2000	2000	2000	2000	48	48	33	33	30	0	0	0	0	1DKLRNEQI
127.436	54.557	43.144	26.021	73.699	27.843	23.064	2000	2000	48	48	33	33	30	29	23	0	0	1KLFRIWY
127.479	54.549	43.144	173.699	27.843	23.064	2000	2000	2000	48	48	33	30	29	23	0	0	0	1DLFRNEQKIWY
129.992	56.625	175.434	43.256	2000	2000	2000	2000	2000	27	27	18	14	0	0	0	0	0	1DKFLEYNQIW
130.777	57.656	180.279	33.289	38.734	183.119	2000	2000	2000	67	67	60	56	37	28	0	0	0	1DQKFLENIWY
130.898	52.869	41.671	174.831	180.743	2000	2000	2000	2000	68	68	52	47	43	0	0	0	0	1DLFNEQKIWY

5.7.3.2. NCOCX List

	67	9																	
	0.5	0.3	0.4	0.3	0.3	0.3	0.3	0.3	0.3										
104.09	54.155	179.389	19.017	2000	2000	2000	2000	2000	67	49	67	57	0	0	0	0	0	1DKTALNIVEQ
105.538	53.467	175.593	38.816	26.011	2000	2000	2000	2000	43	36	43	24	21	0	0	0	0	1DNLEQKI
106.351	59.404	178.926	32.83	173.016	25.726	2000	2000	2000	62	40	62	35	22	16	0	0	0	1MRQVKIDNE
108.096	52.937	174.83	41.712	180.78	2000	2000	2000	2000	61	37	61	26	25	0	0	0	0	1DLNEQKI
108.332	53.657	174.079	40.274	176.885	46.958	173.652	2000	2000	59	47	59	28	25	20	14	0	0	1DLNEQKI
109.015	57.628	177.041	33.251	2000	2000	2000	2000	2000	33	27	33	18	0	0	0	0	0	1MEQKDNLIW
109.445	50.858	176.144	38.275	2000	2000	2000	2000	2000	45	32	45	25	0	0	0	0	0	1DNIAVEQKL
110.615	50.879	176.568	38.274	2000	2000	2000	2000	2000	44	21	44	21	0	0	0	0	0	1DANIYVEQKLFW
112.009	55.718	175.212	28.007	24.357	43.322	2000	2000	2000	49	29	49	28	19	15	0	0	0	1LKI
112.015	56.633	175.584	2000	2000	2000	2000	2000	2000	18	18	18	0	0	0	0	0	0	1HSQKFNREYIVTDLW
113.036	57.179	176.018	39.253	2000	2000	2000	2000	2000	45	15	45	23	0	0	0	0	0	1DFILNYVEQKW
114.161	40.285	176.835	53.633	174.071	2000	2000	2000	2000	47	35	47	31	21	0	0	0	0	1DNLEQKI
114.2	40.285	176.885	53.633	174.071	2000	2000	2000	2000	47	35	47	31	21	0	0	0	0	1DNLEQKI
114.23	40.285	176.947	174.071	2000	2000	2000	2000	2000	35	35	35	21	0	0	0	0	0	1LDNIFVEQKWY
114.889	54.049	174.002	34.266	60.277	2000	2000	2000	2000	64	47	64	34	16	0	0	0	0	1QREVKIDNL
114.909	56.106	175.968	38.298	177.656	2000	2000	2000	2000	64	45	64	30	25	0	0	0	0	1YDNFLEQKIW
115.361	63.722	174.846	31.989	2000	2000	2000	2000	2000	48	43	48	27	0	0	0	0	0	1PVEIDNQK
115.514	2000	179.803	34.057	28.869	2000	2000	2000	2000	78	0	78	19	13	0	0	0	0	1EQKDNLI
116.056	50.889	177.485	23.581	2000	2000	2000	2000	2000	40	32	40	40	0	0	0	0	0	1DPALRNEQKI
116.162	60.471	174.15	72.513	21.82	2000	2000	2000	2000	51	51	51	37	26	0	0	0	0	1ITVK
116.219	55.133	177.067	17.514	2000	2000	2000	2000	2000	44	37	44	44	0	0	0	0	0	1KDLNAEQI
116.283	60.314	171.831	70.565	20.121	174.077	2000	2000	2000	47	47	47	37	26	17	0	0	0	1PITVK
117.238	59.222	177.629	29.011	181.529	35.483	2000	2000	2000	29	29	29	24	20	16	0	0	0	1HQKRWEDNIFY
117.245	54.628	181.208	18.219	2000	2000	2000	2000	2000	44	33	44	44	0	0	0	0	0	1DVPTAEYINQKLFW
118.085	56.143	179.421	18.088	2000	2000	2000	2000	2000	41	38	41	41	0	0	0	0	0	1DTAMILRNVEQK
118.129	60.485	171.987	20.985	71.761	2000	2000	2000	2000	38	36	38	24	20	0	0	0	0	1ITVK
118.544	66.426	178.558	31.839	20.988	2000	2000	2000	2000	33	33	33	25	23	0	0	0	0	1PVKI
118.552	61.044	173.699	73.076	21.158	2000	2000	2000	2000	46	40	46	41	39	0	0	0	0	1ITVK
118.565	66.426	178.578	22.2	31.839	20.988	2000	2000	2000	33	33	33	25	23	21	0	0	0	1PTVKI
118.658	66.426	178.672	31.839	20.988	22.2	2000	2000	2000	53	33	53	25	23	21	0	0	0	1PTVKI
118.885	54.549	177.232	179.611	42.957	2000	2000	2000	2000	40	39	40	38	32	0	0	0	0	1DLANEQKI
118.895	60.323	178.78	32.739	22.2	26.378	2000	2000	2000	53	40	53	30	21	15	0	0	0	1PRVKI
118.957	60.323	178.799	32.739	26.378	2000	2000	2000	2000	53	40	53	30	15	0	0	0	0	1QKPMREDNI
119.221	45.271	172.478	2000	2000	2000	2000	2000	2000	71	64	71	0	0	0	0	0	0	1G
119.496	60.498	175.511	69.901	21.574	2000	2000	2000	2000	50	43	50	42	36	0	0	0	0	1ITVK
120.483	57.626	178.867	37.579	2000	2000	2000	2000	2000	27	18	27	13	0	0	0	0	0	1DKFNILEYVQW
120.512	54.613	179.533	18.313	2000	2000	2000	2000	2000	64	51	64	64	0	0	0	0	0	1DNMLAEQKI
120.76	58.989	177.35	28.862	34.086	179.638	2000	2000	2000	42	38	42	32	25	22	0	0	0	1REQKDNLI
120.816	45.112	171.111	2000	2000	2000	2000	2000	2000	66	56	66	0	0	0	0	0	0	1G
120.912	44.803	172.978	2000	2000	2000	2000	2000	2000	55	45	55	0	0	0	0	0	0	1G
121.033	57.167	179.338	176.015	39.271	27.173	56.12	2000	2000	93	40	93	40	31	16	15	0	0	1DLIRVNEQK
121.034	57.167	179.303	176.015	39.271	56.12	2000	2000	2000	40	40	40	40	31	16	0	0	0	1YDLFNEQKIW
121.045	57.167	179.387	176.015	39.271	60.252	31.674	56.12	27.173	93	40	93	40	31	30	25	16	14	1DFILEYVNQKW
121.045	57.167	179.359	176.015	39.271	60.252	31.674	27.173	56.12	93	40	93	40	31	30	25	16	15	1DFILEYVNQKW
121.051	60.252	179.395	176.015	39.271	31.674	56.12	27.173	2000	93	30	93	40	31	25	16	14	0	1IVKPFREYDNQW
121.629	46.986	173.796	61.859	21.856	2000	2000	2000	2000	68	61	68	15	13	0	0	0	0	1G
122.473	61.538	178.469	38.78	18.09	2000	2000	2000	2000	29	27	29	20	15	0	0	0	0	1EIYFAVDNQKW
122.54	52.646	175	179.743	42.207	2000	2000	2000	2000	34	34	34	32	22	0	0	0	0	1DLNEQKI
122.83	57.144	174.694	43.43	2000	2000	2000	2000	2000	78	32	78	16	0	0	0	0	0	1DFILYVNEQKW
122.996	55.911	174.739	30.335	57.144	43.43	171.192	35.461	2000	78	43	78	33	32	16	15	14	0	1HQKFREYDNLIW
123.074	54.53	173.723	26.056	23.077	27.915	43.231	2000	2000	47	23	47	33	20	20	14	0	0	1LKI
123.079	55.911	174.772	30.335	171.192	35.461	2000	2000	2000	78	43	78	33	15	14	0	0	0	1RHEQKDNLI
123.333	42.746	176.186	51.097	180.225	2000	2000	2000	2000	34	30	34	28	22	0	0	0	0	1DALFNEQKIWY
123.754	67.621	175.657	21.456	54.944	17.521	2000	2000	2000	58	58	58	35	16	15	0	0	0	1DVKTAEINQ
123.993	58.592	172.453	19.728	21.811	32.697	2000	2000	2000	39	33	39	30	30	22	0	0	0	1TVKIL
124.913	54.755	174.836	43.017	27.061	2000	2000	2000	2000	30	27	30	17	15	0	0	0	0	1LRKI
124.947	55.341	177.713	31.568	36.237	2000	2000	2000	2000	38	35	38	27	14	0	0	0	0	1MEQKDNLI
125.178	54.398	171.142	61.491	70.827	32.395	30.205	18.927	174.733	96	59	96	48	44	43	43	37	15	1VQKTMEIDNL
125.466	54.539	175.688	30.874	2000	2000	2000	2000	2000	65	46	65	28	0	0	0	0	0	1KEQRDNLI
126.061	53.053	174.633	27.316	25.073	2000	2000	2000	2000	24	22	24	16	14	0	0	0	0	1LKI
126.256	57.89	171.652	42.721	2000	2000	2000	2000	2000	40	33	40	26	0	0	0	0	0	1IFDLYVNEQKW
126.894	60.283	175.099	37.84	17.59	27.409	59.87	2000	2000	37	37	37	25	24	22	17	0	0	1IVK
126.974	54.918	173.087	36.215	25.582	2000	2000	2000	2000	41	41	41	29	17	0	0	0	0	1QKNMREYDLIFW
127.595	62.065	173.078	69.426	22.569	2000	2000	2000	2000	40	40	40	38	32	0	0	0	0	1TVIK
129.988	62.597	174.147	71.636	21.301	175.325	2000	2000	2000	45	42	45	37	37	15	0	0	0	1PTVIK
130.791	61.409	173.878	72.075	21.458	2000	2000	2000	2000	59	55	59	50	43	0	0	0	0	1PITVK
130.9	61.857	174.958	21.969	31.933	173.758	2000	2000	2000	60	40	60	60	30	17	0	0	0	1PTVIK

5.7.3.3. CANCOCX List

	67	9																	
	0.5	0.4	0.4	0.35	0.35	0.35	0.35	0.35	0.35										
104.096	60.467	179.388	18.999	54.145	2000	2000	2000	2000	27	27	27	27	27	0	0	0	0	1TV
105.543	45.078	175.623	53.438	38.756	26.016	2000	2000	2000	17	17	17	17	13	9	0	0	0	1G
106.382	62.01	178.923	59.335	32.794	2000	2000	2000	2000	19	19	19	19	16	0	0	0	0	1TV
108.037	45.284	174.824	52.897	41.687	2000	2000	2000	2000	17	17	17	17	14	0	0	0	0	1GT
108.05	45.282	174.823	52.897	41.687	2000	2000	2000	2000	17	17	17	17	14	0	0	0	0	1GT
108.271	46.974	174.08	53.648	40.328	2000	2000	2000	2000	21	21	21	21	13	0	0	0	0	1G
108.275	46.965	174.07	53.648	40.328	2000	2000	2000	2000	21	21	21	21	13	0	0	0	0	1G
108.98	61.061	177.052	57.57	33.214	2000	2000	2000	2000	13	13	13	13	9	0	0	0	0	1TVS
109.43	44.846	176.161	50.825	38.236	2000	2000	2000	2000	17	17	17	17	13	0	0	0	0	1G
112.037	62.595	175.201	55.659	28.094	24.409	2000	2000	2000	14	14	14	14	14	9	0	0	0	1SVT
112.134	60.443	175.534	56.6	2000	2000	2000	2000	2000	9	9	9	9	0	0	0	0	0	1CPTVS
114.84	53.672	175.951	56.059	38.306	2000	2000	2000	2000	20	20	20	20	15	0	0	0	0	1DQFELRANKI
114.852	60.263	173.989	54.048	34.261	2000	2000	2000	2000	29	29	29	29	17	0	0	0	0	1ISVKTRYL
115.314	52.614	174.841	63.734	31.989	2000	2000	2000	2000	20	20	20	20	14	0	0	0	0	1LDNQREKI
116.018	63.694	177.475	23.541	50.842	2000	2000	2000	2000	22	22	22	22	17	0	0	0	0	1CISVTKL
116.098	61.426	174.155	60.441	72.514	21.841	2000	2000	2000	26	26	26	26	18	16	0	0	0	1SVTFEYDNQ
116.212	59.229	177.078	55.104	17.488	2000	2000	2000	2000	22	22	22	22	19	0	0	0	0	1HSQKTRVEDNLI
116.271	60.451	171.85	70.522	20.057	2000	2000	2000	2000	26	26	26	20	11	0	0	0	0	1CHSVKQTEDNLI
117.167	67.545	181.187	18.212	54.651	2000	2000	2000	2000	33	33	33	33	32	0	0	0	0	1ITVSKL
117.174	60.31	177.633	59.2	29.032	35.537	2000	2000	2000	17	17	17	17	13	11	0	0	0	1HSVKTFREDNQLI
118.09	57.145	179.423	18.045	56.108	2000	2000	2000	2000	24	24	24	24	22	0	0	0	0	1DSQKMLEYNI
118.229	58.583	172.024	60.451	21.008	71.718	2000	2000	2000	22	22	22	22	16	15	0	0	0	1CSQKMLRVEYDNI
118.502	57.616	178.553	66.394	31.883	22.213	21.037	2000	2000	12	12	12	12	11	9	9	0	0	1DSQKLREYNIF
118.566	57.852	173.693	61.053	21.192	73.065	2000	2000	2000	18	18	18	18	14	13	0	0	0	1DSQKMLRENIY
118.951	66.409	178.799	60.325	32.675	26.369	2000	2000	2000	22	22	22	22	18	10	0	0	0	1ISVKTFEDNQL
118.952	54.11	177.244	54.642	42.99	2000	2000	2000	2000	23	23	23	23	19	0	0	0	0	1DQKNLRAEI
118.953	54.109	177.242	54.642	42.99	2000	2000	2000	2000	23	23	23	23	19	0	0	0	0	1DQKNLRAEI
119.309	55.347	172.5	45.266	2000	2000	2000	2000	2000	39	39	39	39	0	0	0	0	0	1DQKAHLENI
119.543	55.612	175.519	60.48	69.866	21.591	2000	2000	2000	21	21	21	21	19	19	0	0	0	1HQKNAMLEDI
120.437	54.636	179.531	18.265	2000	2000	2000	2000	2000	35	35	35	35	0	0	0	0	0	1DQKALRENI
120.581	60.209	178.959	57.581	2000	2000	2000	2000	2000	12	12	12	12	0	0	0	0	0	1IQVKFHREDNL
120.684	61.577	177.34	58.951	28.835	34.047	2000	2000	2000	16	16	16	16	13	12	0	0	0	1IVTFWYKL
120.747	54.103	171.099	45.106	2000	2000	2000	2000	2000	34	34	34	34	0	0	0	0	0	1DKNLAEQI
120.898	59.386	173.022	44.794	2000	2000	2000	2000	2000	31	31	31	31	0	0	0	0	0	1IVQKLREYDN
120.906	59.385	173.021	44.794	2000	2000	2000	2000	2000	31	31	31	31	0	0	0	0	0	1IVQKLREYDN
121.026	56.099	179.353	57.133	39.204	2000	2000	2000	2000	18	18	18	18	13	0	0	0	0	1DQKFALRENI
121.071	58.946	179.448	60.191	31.673	27.174	2000	2000	2000	18	18	18	18	13	9	0	0	0	1IVQKLREYDN
121.591	61.901	173.789	46.932	46.572	21.901	2000	2000	2000	38	38	38	38	9	9	0	0	0	1IVTFWYKL
122.439	56.148	178.504	61.501	38.763	2000	2000	2000	2000	15	15	15	15	12	0	0	0	0	1DQKFAMLRENI
122.449	54.586	175.013	52.628	42.278	2000	2000	2000	2000	23	23	23	23	17	0	0	0	0	1DQKALENI
122.451	54.588	175.012	52.628	42.278	2000	2000	2000	2000	23	23	23	23	17	0	0	0	0	1DQKAMLENI
122.729	54.762	174.7	43.246	2000	2000	2000	2000	2000	9	9	9	9	0	0	0	0	0	1DKALRENQI
122.732	54.96	174.6	57.119	2000	2000	2000	2000	2000	14	14	14	14	0	0	0	0	0	1DKALRENQI
123.003	53.489	173.768	54.491	43.196	27.878	2000	2000	2000	15	15	15	15	13	10	0	0	0	1DKALRENQI
123.007	53.479	173.77	54.491	43.196	27.878	2000	2000	2000	15	15	15	15	13	10	0	0	0	1DKALRENQI
123.084	57.138	174.76	55.872	30.363	2000	2000	2000	2000	20	20	20	20	16	0	0	0	0	1DSKFILEYNQ
123.392	54.687	176.147	42.728	51.047	2000	2000	2000	2000	16	16	16	16	16	0	0	0	0	1DQKALRENI
123.427	54.683	176.147	51.047	42.728	2000	2000	2000	2000	16	16	16	16	16	0	0	0	0	1DQKALRENI
123.7	55.09	175.651	67.627	21.439	67.229	2000	2000	2000	30	30	30	30	21	16	0	0	0	1QKALREDNI
124.043	61.446	172.481	58.585	21.735	19.754	32.667	2000	2000	16	16	16	16	15	12	11	0	0	1ISVKTFYL
124.876	50.867	174.848	54.802	2000	2000	2000	2000	2000	13	13	13	13	0	0	0	0	0	1CDAHLNEQKI
124.959	57.611	177.677	55.317	31.579	2000	2000	2000	2000	17	17	17	17	13	0	0	0	0	1CDSQKFILEYNW
125.043	55.924	171.227	54.34	32.371	30.175	53.83	2000	2000	29	29	29	29	22	21	8	0	0	1QKFALREYDNI
125.149	50.807	174.82	27.032	2000	2000	2000	2000	2000	9	9	9	9	0	0	0	0	0	1DQPALNEKI
125.179	54.467	171.129	61.44	70.809	18.985	2000	2000	2000	24	24	24	24	21	18	0	0	0	1DKAMHLRENQI
125.37	50.866	175.662	54.53	30.896	2000	2000	2000	2000	26	26	26	26	16	0	0	0	0	1DQPAHLNEKI
126.041	60.112	174.655	53.033	25.78	25.022	42.645	2000	2000	14	14	14	14	9	9	9	0	0	1CIQVKWEDNL
126.073	60.112	174.675	25.78	25.022	53.033	27.28	42.645	2000	14	14	14	14	10	9	9	9	0	1CIVQKEDNL
126.14	60.126	174.714	27.28	25.78	25.022	2000	2000	2000	10	10	10	10	9	9	0	0	0	1CIVKWEDNQL
126.21	51.108	171.68	57.848	42.719	2000	2000	2000	2000	13	13	13	13	11	0	0	0	0	1LDNFAEQKI
126.221	51.1	171.68	57.848	42.719	2000	2000	2000	2000	13	13	13	13	11	0	0	0	0	1LDNFAEQKI
126.89	54.782	175.083	60.18	27.364	37.883	17.549	2000	2000	17	17	17	17	15	12	12	0	0	1DQKALRENI
126.982	53.072	173.098	54.923	36.19	25.541	2000	2000	2000	18	18	18	18	16	11	0	0	0	1DKALRNEQI
127.443	54.52	173.092	61.992	22.613	69.444	2000	2000	2000	20	20	20	20	16	14	0	0	0	1DKELRANQI
129.947	56.638	174.155	62.581	71.624	21.341	2000	2000	2000	18	18	18	18	17	15	0	0	0	1CKFLRWEYDNQI
130.798	52.891	174.941	21.908	61.875	31.954	2000	2000	2000	26	26	26	26	19	13	0	0	0	1HDLRANEQKI
130.799	57.669	173.872	61.395	72.041	21.458	2000	2000	2000	25	25	25	25	23	22	0	0	0	1DVKFHLRWENQI

5.7.4. Example input control file

```
NSGA2_MC_ASSIGN input
!*******************************************************
GB1_seq.txt        !sequence file
3                  !number of spectra
NCACX_50ms.txt     !spectrum 1
NCOCX_50ms.txt     !spectrum 2
CANCOCX_50ms_4D.txt !spectrum 3
Conn_ncacx_ncocx_cancocx_expanded.txt !connection table
NULL          !initial data file
outdata.txt   !output file(data)
outtab        !output files(table)
!*******************************************************
50     !group size
40     !gene pool size
15     !number of steps
5000   !number of NSGA-ii attempts
500000 !number of MC attempts
15     !number of free steps
0.8    !mutation rate
0.6    !additional mutation rate
0.2    !crossover rate
0.3    !null probability
0.0    !minimum w1 (good, for MC step)
0.0    !minimum w2 (bad, for MC step)
0.0    !minimum w3 (edge, for MC step)
0.0    !minimum w4 (used, for MC step)
0.0    !minimum w5 (PIOS weight for MC step)
2.5    !maximum w1 (good, for MC step)
0.0    !maximum w2 (bad, for MC step)
0.5    !maximum w3 (edge, for MC step)
0.5    !maximum w4 (used, for MC step)
1.25   !maximum w5 (PIOS weight for MC step)
!*******************************************************
```

5.7.5. Example connection file for (1) NCACX, (2) NCOCX, and (3) CANcoCX spin-system lists

```
153
  1   2   1   1   0   1     1   3   6   3   0  -1
  1   3   1   1   0   0     1   3   6   4   0  -1
  2   3   1   1   1   0     1   3   6   5   0  -1
  1   2   2   2   0   0     1   3   6   6   0  -1
  1   3   2   2   0   0     1   3   6   7   0  -1
  2   3   2   2   0   0     1   3   6   8   0  -1
  1   2   3   3   0   0     1   3   6   9   0  -1
  1   2   3   4   0   0     1   3   7   3   0  -1
  1   2   3   5   0   0     1   3   7   4   0  -1
  1   2   3   6   0   0     1   3   7   5   0  -1
  1   2   3   7   0   0     1   3   7   6   0  -1
  1   2   3   8   0   0     1   3   7   7   0  -1
  1   2   3   9   0   0     1   3   7   8   0  -1
  1   2   4   3   0   0     1   3   7   9   0  -1
  1   2   4   4   0   0     1   3   8   3   0  -1
  1   2   4   5   0   0     1   3   8   4   0  -1
  1   2   4   6   0   0     1   3   8   5   0  -1
  1   2   4   7   0   0     1   3   8   6   0  -1
  1   2   4   8   0   0     1   3   8   7   0  -1
  1   2   4   9   0   0     1   3   8   8   0  -1
  1   2   5   3   0   0     1   3   8   9   0  -1
  1   2   5   4   0   0     1   3   9   3   0  -1
  1   2   5   5   0   0     1   3   9   4   0  -1
  1   2   5   6   0   0     1   3   9   5   0  -1
  1   2   5   7   0   0     1   3   9   6   0  -1
  1   2   5   8   0   0     1   3   9   7   0  -1
  1   2   5   9   0   0     1   3   9   8   0  -1
  1   2   6   3   0   0     1   3   9   9   0  -1
  1   2   6   4   0   0     2   3   3   3   0  -1
  1   2   6   5   0   0     2   3   3   4   0  -1
  1   2   6   6   0   0     2   3   3   5   0  -1
  1   2   6   7   0   0     2   3   3   6   0  -1
  1   2   6   8   0   0     2   3   3   7   0  -1
  1   2   6   9   0   0     2   3   3   8   0  -1
  1   2   7   3   0   0     2   3   3   9   0  -1
  1   2   7   4   0   0     2   3   4   3   0  -1
  1   2   7   5   0   0     2   3   4   4   0  -1
  1   2   7   6   0   0     2   3   4   5   0  -1
  1   2   7   7   0   0     2   3   4   6   0  -1
  1   2   7   8   0   0     2   3   4   7   0  -1
  1   2   7   9   0   0     2   3   4   8   0  -1
  1   2   8   3   0   0     2   3   4   9   0  -1
  1   2   8   4   0   0     2   3   5   3   0  -1
  1   2   8   5   0   0     2   3   5   4   0  -1
  1   2   8   6   0   0     2   3   5   5   0  -1
  1   2   8   7   0   0     2   3   5   6   0  -1
  1   2   8   8   0   0     2   3   5   7   0  -1
  1   2   8   9   0   0     2   3   5   8   0  -1
  1   2   9   3   0   0     2   3   5   9   0  -1
  1   2   9   4   0   0     2   3   6   3   0  -1
  1   2   9   5   0   0     2   3   6   4   0  -1
  1   2   9   6   0   0     2   3   6   5   0  -1
  1   2   9   7   0   0     2   3   6   6   0  -1
  1   2   9   8   0   0     2   3   6   7   0  -1
  1   2   9   9   0   0     2   3   6   8   0  -1
  1   3   3   3   0  -1     2   3   6   9   0  -1
  1   3   3   4   0  -1     2   3   7   3   0  -1
  1   3   3   5   0  -1     2   3   7   4   0  -1
  1   3   3   6   0  -1     2   3   7   5   0  -1
  1   3   3   7   0  -1     2   3   7   6   0  -1
  1   3   3   8   0  -1     2   3   7   7   0  -1
  1   3   3   9   0  -1     2   3   7   8   0  -1
  1   3   4   3   0  -1     2   3   7   9   0  -1
  1   3   4   4   0  -1     2   3   8   3   0  -1
  1   3   4   5   0  -1     2   3   8   4   0  -1
  1   3   4   6   0  -1     2   3   8   5   0  -1
  1   3   4   7   0  -1     2   3   8   6   0  -1
  1   3   4   8   0  -1     2   3   8   7   0  -1
  1   3   4   9   0  -1     2   3   8   8   0  -1
  1   3   5   3   0  -1     2   3   8   9   0  -1
  1   3   5   4   0  -1     2   3   9   3   0  -1
  1   3   5   5   0  -1     2   3   9   4   0  -1
  1   3   5   6   0  -1     2   3   9   5   0  -1
  1   3   5   7   0  -1     2   3   9   6   0  -1
  1   3   5   8   0  -1     2   3   9   7   0  -1
  1   3   5   9   0  -1     2   3   9   8   0  -1
                            2   3   9   9   0  -1
```

5.8. REFERENCES

(1) Comellas, G.; Rienstra, C. M. Protein structure determination by magic-angle spinning solid-state NMR, and insights into the formation, structure, and stability of amyloid fibrils. *Annu. Rev. Biophys.* **2013**, *42*, 515–536.

(2) Spera, S.; Bax, A. Empirical correlation between protein backbone conformation and Cα and Cβ 13C nuclear magnetic resonance chemical shifts. *J. Am. Chem. Soc.* **1991**, *113*, 5490–5492.

(3) Zhang, H. Y.; Neal, S.; Wishart, D. S. RefDB: A database of uniformly referenced protein chemical shifts. *J. Biomol. NMR* **2003**, *25*, 173–195.

(4) Gullion, T.; Schaefer, J. Rotational-echo double-resonance NMR. *J. Magn. Reson.* **1989**, *81*, 196–200.

(5) De Paepe, G.; Lewandowski, J. R.; Loquet, A.; Bockmann, A.; Griffin, R. G. Proton assisted recoupling and protein structure determination. *J. Chem. Phys.* **2008**, *129*, 245101.

(6) Rienstra, C. M.; Hohwy, M.; Mueller, L. J.; Jaroniec, C. P.; Reif, B.; Griffin, R. G. Determination of multiple torsion-angle constraints in U–13C,15N-labeled peptides: 3D 1H–15N–13C–1H dipolar chemical shift NMR spectroscopy in rotating solids. *J. Am. Chem. Soc.* **2002**, *124*, 11908–11922.

(7) Shi, X.; Rienstra, C. M. Site-Specific Internal Motions in GB1 Protein Microcrystals Revealed by 3D (2)H-(1)(3)C-(1)(3)C Solid-State NMR Spectroscopy. *J. Am. Chem. Soc.* **2016**, *138*, 4105–4119.

(8) Wylie, B. J.; Bhate, M. P.; McDermott, A. E. Transmembrane allosteric coupling of the gates in a potassium channel. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 185–190.

(9) Schwieters, C. D.; Kuszewski, J. J.; Tjandra, N.; Clore, G. M. The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **2003**, *160*, 65–73.

(10) Rienstra, C. M.; Hohwy, M.; Hong, M.; Griffin, R. G. 2D and 3D 15N-13C-13C NMR chemical shift correlation spectroscopy of solids: Assignment of MAS spectra of peptides. *J. Am. Chem. Soc.* **2000**, *122*, 10979–10990.

(11) Baldus, M.; Petkova, A. T.; Herzfeld, J.; Griffin, R. G. Cross polarization in the tilted frame: assignment and spectral simplification in heteronuclear spin systems. *Mol. Phys.* **1998**, *95*, 1197–1207.

(12) Franks, W. T.; Kloepper, K. D.; Wylie, B. J.; Rienstra, C. M. Four-dimensional heteronuclear correlation experiments for chemical shift assignment of solid proteins. *J. Biomol. NMR* **2007**, *39*, 107–131.

(13) Linser, R.; Bardiaux, B.; Andreas, L. B.; Hyberts, S. G.; Morris, V. K.; Pintacuda, G.; Sunde, M.; Kwan, A. H.; Wagner, G. Solid-state NMR structure determination from diagonal-compensated, sparsely nonuniform-sampled 4D proton-proton restraints. *J. Am. Chem. Soc.* **2014**, *136*, 11002–11010.

(14) Hoch, J. C.; Maciejewski, M. W.; Mobli, M.; Schuyler, A. D.; Stern, A. S. Nonuniform sampling and maximum entropy reconstruction in multidimensional NMR. *Acc. Chem. Res.* **2014**, *47*, 708–717.

(15) Palmer, M. R.; Suiter, C. L.; Henry, G. E.; Rovnyak, J.; Hoch, J. C.; Polenova, T.; Rovnyak, D. Sensitivity of nonuniform sampling NMR. *J. Phys. Chem. B* **2015**, *119*, 6502–6515.

(16) Ying, J.; Delaglio, F.; Torchia, D. A.; Bax, A. Sparse multidimensional iterative lineshape-enhanced (SMILE) reconstruction of both non-uniformly sampled and conventional NMR data. *J. Biomol. NMR* **2017**, *68*, 101–118.

(17) Goddard, T. D.; Kneller, D. G. SPARKY 3. University of California, San Francisco, **2008**.

(18) Lee, W.; Tonelli, M.; Markley, J. L. NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* **2015**, *31*, 1325–1327.

(19) Tycko, R.; Hu, K. N. A Monte Carlo/simulated annealing algorithm for sequential resonance assignment in solid state NMR of uniformly labeled proteins with magic-angle spinning. *J. Magn. Reson.* **2010**, *205*, 304–314.

(20) Hu, K. N.; Qiang, W.; Tycko, R. A general Monte Carlo/simulated annealing algorithm for resonance assignment in NMR of uniformly labeled biopolymers. *J. Biomol. NMR* **2011**, *50*, 267–276.

(21) Yang, Y.; Fritzsching, K. J.; Hong, M. Resonance assignment of the NMR spectra of disordered proteins using a multi-objective non-dominated sorting genetic algorithm. *J. Biomol. NMR* **2013**, *57*, 281–296.

(22) Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **2002**, *6*, 182–197.

(23) Murray, D. T.; Kato, M.; Lin, Y.; Thurber, K. R.; Hung, I.; McKnight, S. L.; Tycko, R. Structure of FUS protein fibrils and its relevance to self-assembly and phase separation of low-complexity domains. *Cell* **2017**, *171*, 615–627 e616.

(24) Stevens, T. J.; Fogh, R. H.; Boucher, W.; Higman, V. A.; Eisenmenger, F.; Bardiaux, B.; van Rossum, B. J.; Oschkinat, H.; Laue, E. D. A software framework for analysing solid-state MAS NMR data. *J. Biomol. NMR* **2011**, *51*, 437–447.

(25) Johnson, B. A. Using NMRView to visualize and analyze the NMR spectra of macromolecules. *Methods Mol Biol* **2004**, *278*, 313–352.

(26) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. NMRPipe: a multidimensional spectral processing system based on unix pipes. *J. Biomol. NMR* **1995**, *6*, 277–293.

(27) Bahrami, A.; Assadi, A. H.; Markley, J. L.; Eghbalnia, H. R. Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein NMR spectroscopy. *PLoS Comput Biol* **2009**, *5*, e1000307.

(28) Fritzsching, K. J.; Yang, Y.; Schmidt-Rohr, K.; Hong, M. Practical use of chemical shift databases for protein solid-state NMR: 2D chemical shift maps and amino-acid assignment with secondary-structure information. *J. Biomol. NMR* **2013**, *56*, 155–167.

(29) Piehl, D. W.; Blancas-Mejia, L. M.; Ramirez-Alvarado, M.; Rienstra, C. M. Solid-state NMR chemical shift assignments for AL-09 VL immunoglobulin light chain fibrils. *Biomol. NMR Assign.* **2017**, *11*, 45–50.

(30) Piehl, D. W.; Blancas-Mejia, L. M.; Wall, J. S.; Kennel, S. J.; Ramirez-Alvarado, M.; Rienstra, C. M. Immunoglobulin light chains form an extensive and highly ordered fibril involving the N- and C-termini. *ACS Omega* **2017**, *2*, 712–720.

CHAPTER 6: CONCLUSIONS AND OUTLOOK

6.1. CONCLUSIONS

In this work, I presented the collection of investigations we have performed that are directed at understanding the structural determinants of cardiac light-chain amyloidosis, through the application of MAS SSNMR spectroscopy to the case of AL-09 V_L fibrils. Overall, our results demonstrate that the majority of the 108 amino acid sequence of AL-09 V_L is highly rigid in the fibril structure and that the N- and C- termini in particular exhibit a substantial extent of order relative to the other regions of the protein sequence[1,2]. Furthermore, recent insights described in Chapter 4 reveal that several of the centrally-occurring residues in the sequence also exhibit a moderate degree of order in addition to a set of distinct chemical shift frequencies that may indicate the adoption of a unique conformational arrangement relative to the surrounding topology of the fibril core. Collectively, these results and conclusions offer several of the first key insights into the molecular structure and dynamics involved in AL disease fibrils, and importantly place the project at a highly optimal position to pursue the measurement of distance restraints and subsequent simulated annealing calculations for the determination of the first three-dimensional AL fibril structure to be reported.

In light of the multitude of AL protein sequences that can form fibrils between patients, there remains a large urgency for the elucidation of a fibril structure at the molecular-level, as this may serve as a model on which the variety of other AL fibrils may be studied. Indeed, several structurally-based investigations on disease-related amyloid fibrils have yielded a variety of molecules with potentially therapeutic and diagnostic significance, particularly in the case of Alzheimer's disease[3,4]. These studies are enabled by the availability of the three-dimensional structure of the corresponding fibril system[3]. Accordingly, the future determination of the structure of AL fibrils may fuel similar studies. Nonetheless, despite the current absence of a full three-dimensional picture for the AL fibril species, several studies have reported the identification of a murine monoclonal antibody that can specifically bind the fibril form of the protein, but not the soluble form[5,6]. Thus, there exists a great amount of potential for not only extending these findings into diagnostic applications, but also for providing model framework through

which we may probe these fibril-specific binding interactions to inform subsequent research on the rational design of other molecules that may participate in similar binding interactions.

6.2. OUTLOOK

As described above, there remains a large demand for obtaining a complete description of the atomic-resolution structure of AL fibrils. Towards this interest, the results and conclusions provided in the preceding chapters place us at a particularly opportune position for achieving this objective in the case of AL-09 V_L fibrils. Specifically, our knowledge of the chemical shifts for approximately 90% of the structurally-rigid AL-09 V_L protein backbone will enable us to perform the final set of experiments necessary for determining a comprehensive list of distance restraints, which we may subsequently incorporate into simulated annealing structure calculations[7]. Moreover, as we now also have access to an extensive list of backbone torsion angle predictions (as well as for side-chain χ^1 angles) determined from TALOS-N[8], these may be supplied as additional structural restraints into the calculations.

In considering the various experimental strategies that can be pursued to obtain a valuable set of distance restraints, several in particular have been demonstrated to provide success in the case of fibril systems—both through the results of previous studies[9-12] as well as ongoing research efforts in our group—and differ depending on the type of internuclear distance to be measured (i.e., intra- or intermolecular) as well as the labeling pattern of the sample[13]. For the case of uniformly [^{13}C, ^{15}N]-labeled samples, which represents the labeling scheme of all AL-09 V_L fibril preparations thus far, one can employ homonuclear dipolar-recoupling pulse sequence techniques such as proton assisted recoupling (PAR)[14,15], dipolar assisted rotational-recoupling (DARR)[16], and proton-driven spin-diffusion (PDSD)[17] to enable the measurement of short-, medium-, and long-range distances between ^{13}C–^{13}C (PAR, DARR, and PDSD) or ^{15}N–^{15}N (PDSD[18] and PAR[19]) spin pairs. Furthermore, heteronuclear dipolar-recoupling experiments such as proton assisted insensitive nuclei cross polarization (PAIN-CP)[20] or rotational-echo double-resonance (REDOR)[21] can be performed to obtain ^{13}C–^{15}N distance restraints, using U-[^{13}C, ^{15}N] samples that are fully labeled[22,23] or diluted in natural abundance material[24], respectively. Accordingly, in

light of the current availability of solely U-[^{13}C, ^{15}N]-labeled AL-09 V_L fibrils, the application of these techniques (particularly, PAR and PAIN-CP) would provide a direct route to obtaining the first set of intramolecular distances for the fibril structure.

Additionally, the preparation of sparsely ^{13}C-labeled samples offers another powerful approach to the determination of unambiguous medium- and long-range distance restraints for both intra- and intermolecular correlations. While there are a variety of possible sparse-labeling schemes that can be prepared, a few that are more commonly used and likely most valuable for the case of AL-09 V_L fibrils include [1,3-^{13}C]- or [2-^{13}C]-glycerol[25] as well as [1-^{13}C]- or [2-^{13}C]-glucose[26], which are provided as the growth substrate to the cellular culture during protein expression (with or without $^{15}NH_3$, depending on the type of distance desired). The study of these samples through the similar application of the techniques above—namely, REDOR and the related transferred-echo double-resonance (TEDOR)[27]—can enable for a significant improvement to the number of intramolecular restraints obtained from the uniformly-labeled samples alone, as the sparse-labeling reduces spectral congestion. Furthermore, by combining these sparsely ^{13}C-labeled samples with solely ^{15}N-labeled material (prior to fibril formation), one can enable to additional determination of unambiguous intermolecular distances. Collectively, these experiments can be used to obtain a large variety of distance restraints for the fibril sample of interest, and are the most beneficial when all of the chemical shifts for the protein are known, as this allows for the unambiguous identification of the residues corresponding to long-range cross-peak correlations. Thus, the extension of these investigations to the case of AL-09 V_L fibrils should be a relatively straightforward task following the necessary sample preparation, which is currently underway.

The distance restraint information obtained from the above experiments may be further complemented by backbone torsion angle measurements for specific sites along the protein sequence, in order to provide an augmented list of dihedral angle restraints. A variety of experiments have been developed to perform these measurements in the solid state, including the T-MREV[28,29] and BARE[30] pulse sequences. The application of these experiments to AL-09 V_L fibrils would be particularly valuable for investigating the regions of the sequence for which TALOS-N was unable to make accurate predictions of

the corresponding angles, such as the stretch of residues from T56–T69 (as described in Chapter 4). Moreover, the success of this approach has been demonstrated by Murray et al. in the case of FUS-LC fibrils, in which ^{15}N-BARE measurements were performed to obtain restraints for multiple residues within the β-loop structure of the fibril core[12]. Notably, as discussed in Chapter 4, this β-loop structure also exhibits several similarities in sequence and chemical shift as the T56–T69 stretch in AL-09 V_L, and thus represents one of the proposed structural models for this region of the fibril core.

Together, the structural restraints that we obtain from each of these methods can be incorporated into simulated annealing calculations using the XPLOR-NIH software package[7]. Ultimately, the iterative process of performing these calculations and eliminating unreasonable solutions—combined with fine-tuning of the provided set of restraints—may allow for identification of a dominant fold between different conformers and eventually a final high-resolution structure that demonstrates good agreement with the data. Through these means, the three-dimensional molecular structure of AL-09 V_L fibrils will be determined.

6.3. REFERENCES

(1) Piehl, D. W.; Blancas-Mejia, L. M.; Ramirez-Alvarado, M.; Rienstra, C. M. Solid-state NMR chemical shift assignments for AL-09 VL immunoglobulin light chain fibrils. *Biomol. NMR Assign.* **2017**, *11*, 45–50.

(2) Piehl, D. W.; Blancas-Mejia, L. M.; Wall, J. S.; Kennel, S. J.; Ramirez-Alvarado, M.; Rienstra, C. M. Immunoglobulin light chains form an extensive and highly ordered fibril involving the N- and C-termini. *ACS Omega* **2017**, *2*, 712–720.

(3) Jiang, L.; Liu, C.; Leibly, D.; Landau, M.; Zhao, M. L.; Hughes, M. P.; Eisenberg, D. S. Structure-based discovery of fiber-binding compounds that reduce the cytotoxicity of amyloid beta. *eLife* **2013**, *2*, e00857.

(4) Sievers, S. A.; Karanicolas, J.; Chang, H. W.; Zhao, A.; Jiang, L.; Zirafi, O.; Stevens, J. T.; Munch, J.; Baker, D.; Eisenberg, D. Structure-based design of non-natural amino-acid inhibitors of amyloid fibril formation. *Nature* **2011**, *475*, 96–103.

(5) Hrncic, R.; Wall, J.; Wolfenbarger, D. A.; Murphy, C. L.; Schell, M.; Weiss, D. T.; Solomon, A. Antibody-mediated resolution of light chain-associated amyloid deposits. *Am. J. Pathol.* **2000**, *157*, 1239–1246.

(6) O'Nuallain, B.; Allen, A.; Kennel, S. J.; Weiss, D. T.; Solomon, A.; Wall, J. S. Localization of a conformational epitope common to non-native and fibrillar immunoglobulin light chains. *Biochemistry* **2007**, *46*, 1240–1247.

(7) Schwieters, C. D.; Kuszewski, J. J.; Tjandra, N.; Clore, G. M. The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **2003**, *160*, 65–73.

(8) Shen, Y.; Bax, A. Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J. Biomol. NMR* **2013**, *56*, 227–241.

(9) Lu, J. X.; Qiang, W.; Yau, W. M.; Schwieters, C. D.; Meredith, S. C.; Tycko, R. Molecular structure of β-amyloid fibrils in Alzheimer's disease brain tissue. *Cell* **2013**, *154*, 1257–1268.

(10) Tuttle, M. D.; Comellas, G.; Nieuwkoop, A. J.; Covell, D. J.; Berthold, D. A.; Kloepper, K. D.; Courtney, J. M.; Kim, J. K.; Barclay, A. M.; Kendall, A.; Wan, W.; Stubbs, G.; Schwieters, C. D.; Lee, V. M. Y.; George, J. M.; Rienstra, C. M. Solid-state NMR structure of a pathogenic fibril of full-length human α-synuclein. *Nat. Struct. Mol. Biol.* **2016**, *23*, 409–415.

(11) Colvin, M. T.; Silvers, R.; Ni, Q. Z.; Can, T. V.; Sergeyev, I.; Rosay, M.; Donovan, K. J.; Michael, B.; Wall, J.; Linse, S.; Griffin, R. G. Atomic Resolution Structure of Monomorphic Aβ42 Amyloid Fibrils. *J. Am. Chem. Soc.* **2016**, *138*, 9663–9674.

(12) Murray, D. T.; Kato, M.; Lin, Y.; Thurber, K. R.; Hung, I.; McKnight, S. L.; Tycko, R. Structure of FUS protein fibrils and its relevance to self-assembly and phase separation of low-complexity domains. *Cell* **2017**, *171*, 615–627 e616.

(13) Comellas, G.; Rienstra, C. M. Protein structure determination by magic-angle spinning solid-state NMR, and insights into the formation, structure, and stability of amyloid fibrils. *Annu. Rev. Biophys.* **2013**, *42*, 515–536.

(14) De Paepe, G.; Lewandowski, J. R.; Loquet, A.; Bockmann, A.; Griffin, R. G. Proton assisted recoupling and protein structure determination. *J. Chem. Phys.* **2008**, *129*, 245101.

(15) Donovan, K. J.; Jain, S. K.; Silvers, R.; Linse, S.; Griffin, R. G. Proton-Assisted Recoupling (PAR) in Peptides and Proteins. *J. Phys. Chem. B* **2017**, *121*, 10804–10817.

(16) Takegoshi, K.; Nakamura, S.; Terao, T. 13C-1H dipolar-assisted rotational resonance in magic-angle spinning NMR. *Chem. Phys. Lett.* **2001**, *344*, 631–637.

(17) Suter, D.; Ernst, R. R. Spectral Spin Diffusion in the Presence of an Extraneous Dipolar Reservoir. *Phys. Rev. B* **1982**, *25*, 6038–6041.

(18) Franks, W. T.; Wylie, B. J.; Stellfox, S. A.; Rienstra, C. M. Backbone conformational constraints in a microcrystalline U-15N-labeled protein by 3D dipolar-shift solid-state NMR spectroscopy. *J. Am. Chem. Soc.* **2006**, *128*, 3154–3155.

(19) Lewandowski, J. R.; De Paepe, G.; Eddy, M. T.; Griffin, R. G. (15)N-(15)N proton assisted recoupling in magic angle spinning NMR. *J. Am. Chem. Soc.* **2009**, *131*, 5769–5776.

(20) Lewandowski, J. R.; De Paepe, G.; Griffin, R. G. Proton assisted insensitive nuclei cross polarization. *J. Am. Chem. Soc.* **2007**, *129*, 728–729.

(21) Gullion, T.; Schaefer, J. Rotational-echo double-resonance NMR. *J. Magn. Reson.* **1989**, *81*, 196–200.

(22) Bertini, I.; Bhaumik, A.; De Paepe, G.; Griffin, R. G.; Lelli, M.; Lewandowski, J. R.; Luchinat, C. High-resolution solid-state NMR structure of a 17.6 kDa protein. *J. Am. Chem. Soc.* **2010**, *132*, 1032–1040.

(23) Nielsen, A. B.; Szekely, K.; Gath, J.; Ernst, M.; Nielsen, N. C.; Meier, B. H. Simultaneous acquisition of PAR and PAIN spectra. *J. Biomol. NMR* **2012**, *52*, 283–288.

(24) Rienstra, C. M.; Tucker-Kellogg, L.; Jaroniec, C. P.; Hohwy, M.; Reif, B.; McMahon, M. T.; Tidor, B.; Lozano-Perez, T.; Griffin, R. G. De novo determination of peptide structure with solid-state magic-angle spinning NMR spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 10260–10265.

(25) Higman, V. A.; Flinders, J.; Hiller, M.; Jehle, S.; Markovic, S.; Fiedler, S.; van Rossum, B. J.; Oschkinat, H. Assigning large proteins in the solid state: a MAS NMR resonance assignment strategy using selectively and extensively 13C-labelled proteins. *J. Biomol. NMR* **2009**, *44*, 245–260.

(26) Hong, M. Determination of multiple phi-torsion angles in proteins by selective and extensive 13C labeling and two-dimensional solid-state NMR. *J. Magn. Reson.* **1999**, *139*, 389–401.

(27) Hing, A. W.; Vega, S.; Schaefer, J. Transferred-Echo Double-Resonance NMR. *J. Magn. Reson.* **1992**, *96*, 205–209.

(28) Hohwy, M.; Jaroniec, C. P.; Reif, B.; Rienstra, C. M.; Griffin, R. G. Local structure and relaxation in solid-state NMR: accurate measurement of amide N-H bond lengths and H-N-H bond angles. *J. Am. Chem. Soc.* **2000**, *122*, 3218–3219.

(29) Rienstra, C. M.; Hohwy, M.; Mueller, L. J.; Jaroniec, C. P.; Reif, B.; Griffin, R. G. Determination of multiple torsion-angle constraints in U–13C,15N-labeled peptides: 3D 1H–15N–13C–1H dipolar chemical shift NMR spectroscopy in rotating solids. *J. Am. Chem. Soc.* **2002**, *124*, 11908–11922.

(30) Hu, K. N.; Qiang, W.; Bermejo, G. A.; Schwieters, C. D.; Tycko, R. Restraints on backbone conformations in solid state NMR studies of uniformly labeled proteins from quantitative amide 15N-15N and carbonyl 13C-13C dipolar recoupling data. *J. Magn. Reson.* **2012**, *218*, 115–127.