Modeling Intrinsically Disordered Proteins;
A Comprehensive Study of α-Synuclein

By

Orly Ullman

B.S. Chemistry
Tel-Aviv University, Tel-Aviv, Israel, 2005

Submitted to the Department of Chemistry
In Partial Fulfillment of the Requirement for the Degree of

Doctor of Philosophy
in Physical Chemistry

at the

Massachusetts Institute of Technology

February 2015

Signature of Author: Signature redacted
Department of Chemistry
January 15, 2015

Certified by: Signature redacted
Collin M. Stultz
Professor of Electrical Engineering and Computer Science
and the Institute for Medical Engineering and Science
Thesis Supervisor

Accepted by: Signature redacted
Robert W. Field
Chairman, Departmental Committee on Graduate Students

This doctoral thesis has been examined by a Committee of the Department of Chemistry as follows:

Professor Jianshu Cao Signature redacted

Committee Chairman
Professor of Chemistry

Professor Collin M. Stultz Signature redacted

Research Supervisor
Professor of Electrical Engineering and Computer Science
and the Institute for Medical Engineering and Science

Professor Catherine L. Drennan Signature redacted

Committee member
Professor of Chemistry and Biology
Howard Hughes Medical Investigator and Professor

Modeling Intrinsically Disordered Proteins;
A Comprehensive Study of α-Synuclein

by Orly Ullman

Submitted to the Department of Chemistry in January 2015
In Partial Fulfillment of the Requirement for the Degree of
Doctor of Philosophy in Physical Chemistry

Thesis Supervisor: Collin M. Stultz
Title: Professor of Electrical Engineering and Computer Science and the Institute for Medical Engineering and Science

ABSTRACT

Parkinson's disease (PD) affects over 10 million people worldwide and has no cure. Moreover, current treatments for PD have limited efficacy. Studies that advance our understanding of the mechanism of neurodegeneration in PD will provide guidance in our search for effective therapies for this neurodegenerative disorder.

PD is characterized clinically by motor deficits – namely resting tremors, rigidity, bradykinesia and postural instability – and pathologically by intraneuronal inclusions in the substantia nigra. Several studies suggest that α-synuclein, the major component of these intracellular inclusions, plays a major role in the neurodegenerative process. Therefore understanding the structural properties of α-synuclein and its aggregation mechanism is of particular interest.

α-synuclein is particularly challenging to study because it is an Intrinsically Disordered Protein (IDP); i.e., it lacks a well-defined structure in aqueous solution. Unlike folded proteins, IDPs typically interconvert between many different conformations during their biological lifetime. In this thesis we apply novel methods to develop models for IDPs and apply them to α-synuclein. The overriding hypothesis that forms the basis of this work is that IDPs in solution can be modeled as a finite set of energetically favorable structures, where each structure corresponds to an energy minimum on a complex energy landscape. The number of structures in the resulting ensemble is related to the resolution in which one wishes to view the energy landscape of the protein. We demonstrate that this approach leads to new insights into the aggregation mechanism of α-synuclein.

Acknowledgments

There's a phrase in Hebrew which goes "קצרה היריעה מלהכיל". In a somewhat loose translation it means "The space is insufficient to contain". I feel this phrase is a good description of my feelings in writing these acknowledgments. Many people contributed to my growth during the constructive years I spent as a graduate student in MIT. I will not be able to thank them all, and certainly will not be able to do justice with these limited words of gratitude. Therefore the following acknowledgments are ill-defined and incomplete. Luckily, as the sentimental part of this document they can suffer from such ills.

First, I'd like to thank the members of my thesis committee; Professor Cathy Drennan and Professor Jianshu Cao. In addition I'd like to pay my respects and gratitude to the late Professor Bob Silbey who served as my previous thesis committee chair. These esteemed faculty members helped shape this thesis with their thoughtful comments.

The main figure I wish to thank in writing this dissertation is my thesis supervisor, Professor Collin Stultz. Over the years I learned a great deal from Collin about how to be a scientist and how to conduct research. He has been a lead influence in shaping my critical thinking skills and accommodated some of my more eccentric endeavors. In an age where one spends most of their time in the work place, it is best for the work place to feel as a second home. Collin has a great gift for generating that feeling: The people he joins together to form the Stultz lab have always been a collection of creative, smart and kind individuals who make going to work something to look forward to. I view them all as family.

I thank the lab mates who formed this extended home. Dr. Ausin Huang, who helped me settle in and dive head first into CHARMM, shell scripting, clusters and all the wonderful other programming concepts that quickly became the basic toolset of my computational work. Dr. Sophie Walker, our then resident experimentalist, who made working in a wet-lab seem fun and interesting, and became a dear friend as well as a scientific collaborator. Dr. Charles Fisher who I was lucky to overlap and work with closely over his entire graduate career in the lab, we collaborated on many projects and he always had a talent for posing challenging questions that made me think and rethink things along the way, I'm lucky have him as a friend and colleague. Dr. Thomas Gurry, another wonderful soul, who I also overlapped with for several years and collaborated with in writing the last chapter of this dissertation. I've had the opportunity to

interact with many more amazing people in the Stultz lab: Dr. Chris Schubert, Dr. Paul Nerenberg, Dr. Elaine Gee, Dr. Ramon Salsas-Escat, Dr. Christine Phillips-Piro, Dr. Sarah Bowman, Munishika Kalia, Dr. Virginia Burger, Dr. Linder Candido da Silva, Yun Liu, Dr. Diego Nolasco, Priya Parayanthal, Mathura Sridharan and the soon-to-be medical doctors Gordon Lu and Joyatee Sarker. With all these people I shared wonderful intellectual and not-so-intellectual much needed breaks and conversations throughout the years.

Although they did not have a direct hand in writing this thesis, I'd like to add a special thank you to Professor John Guttag and the wonderful members of his lab: Dr. Jenna Weins, Dr. Garthee Ganeshpapillai, Anima Singh, Joel Brooks, Guha Balakrishnan, Jen Gong, and Amy Zhao. During my PhD at MIT, I became interested in Machine Learning and was kindly adopted by the Guttag lab, they introduced me to a completely different language with an astounding patience and willingness to teach.

To my family, genetic and lawful, those in faraway lands and those that are always near, T & g, you are my heart and hearth.

Table of Contents

List of Figures

Chapter I

Studying Intrinsically Disordered Proteins Using Computational Methods

I.A Introduction

The archetypical structure-function paradigm states that there exists a unique mapping between a protein's function and its native structure. However, a growing body of work suggests that many proteins do not adopt a single folded state under physiological conditions (Uversky 2002, Radivojac, Iakoucheva et al. 2007). Some proteins, such as λ-repressor, have short disordered linkers between two separate, folded, domains (Zhou 2001). Others, including a variety of other transcription factors such as p53, contain folded domains associated with extended regions of disorder, which are thought to facilitate the recognition of different protein targets (von Hippel 2004, Liu, Perumal et al. 2006).

The fact that many proteins sample a heterogeneous set of conformations during their biological lifetime argues for an alternate view of protein structure. In this regard, classification schemes based on an order-disorder continuum are likely to be more à propos. On one end of the spectrum are proteins that adopt a unique fold under physiological conditions and on the other end are proteins classified as *intrinsically disordered proteins* (IDPs) (Fisher and Stultz 2011) which sample a large number of structurally dissimilar states in solution (Huang and Stultz 2009). As a result, traditional methods for protein structure determination that employ NMR spectroscopy or X-ray crystallography, which while appropriate for studying folded proteins, are ill suited for studying IDPs.

Despite these difficulties studying IDPs is of paramount importance because many have been implicated as key protagonists in a number of neurodegenerative disorders, including Alzheimer's, Parkinson's and Huntington's disease, through their tendency to aggregate into toxic structures (Huang and Stultz 2009). As such, they constitute a sought after starting-point for rational drug design methods aimed at alleviating, or reversing, the symptoms of these disorders. In order to move forward it is therefore critical to understand the thermally accessible

state of IDPs. Several computational methods have been developed to provide insight into the underlying set of accessible conformations that IDP can adopt during its biological lifetime.

I.A.1 What are intrinsically disordered proteins?

It has been suggested that the energy landscape of a natively folded protein resembles a funnel, where the minimum corresponds to the most thermodynamically stable conformation. The global minimum is structurally unique and the fact that such proteins remain folded for much of their biological lifetime suggests that the free energy minimum is significantly lower than any other state. The free energy landscape of an IDP is, by contrast, comparatively 'flat' (see Figure I.1). The shape of the energy landscape enables the protein to sample a large number of local energy minima through thermal fluctuations alone (Uversky, Gillespie et al. 2000).

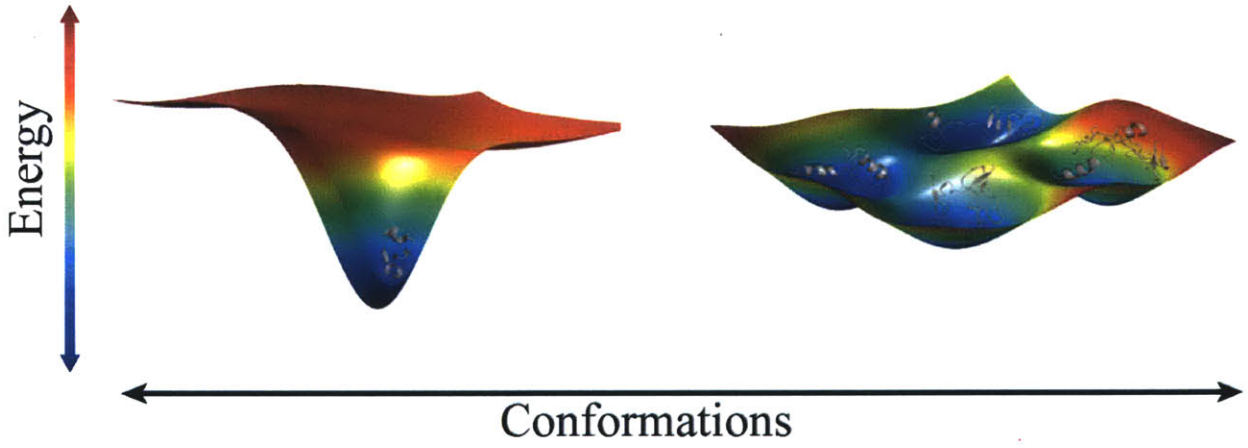

Figure I.1: The 'funnel' shaped energy landscape for natively folded proteins (on the left) in comparison to the relatively 'flat' energy landscape of an IDP (on the left).

To describe an IDP's structure, one must account for this inherent conformational heterogeneity. In this thesis, we demonstrate how computational methods can be used to understand the thermodynamics of IDPs. Our overriding hypothesis is that IDPs in solution can be modeled as a finite set of energetically favorable structures where each structure corresponds to an energy minimum over a complex energy landscape. We illustrate these principles using the IDP α-synuclein.

I.B Generating Ensembles for Intrinsically Disordered Proteins

IDPs can be modeled using structural ensembles. An ensemble is defined as a set of N structures $S_N = \{s_1, s_2, \ldots, s_N\}$, (a structural library) and a set of their relative stabilities or weights $W_N = \{w_1, w_2, \ldots, w_N\}$ (Huang and Stultz 2008, Huang and Stultz 2009, Fisher, Huang et al. 2010, Fisher and Stultz 2011, Fisher and Stultz 2011). By choosing N, the number of structures, one is effectively choosing the resolution in which one view's the IDP's energy landscape. In practice, structures and weights are generated such that calculated ensemble average quantities agree with a pre-specified set of experimental observables.

The first step in these ensemble construction methods is the generation of a diverse structural library; this is typically done using MD simulations, Monte Carlo sampling methods, or statistical coil models. As the sizes of the proteins that are of interest are relatively large, special methods are needed to ensure that a wide range of conformational space is sampled in a reasonable amount of CPU time. Methods such as replica exchange (Sugita and Okamoto 1999), accelerated MD (Chen and Horing 2007) or fragmenting the protein using MD on each of the segments followed by piecing the segments together (Fisher, Huang et al. 2010) all facilitate a broad sampling of conformational space. Statistical coil models use a residue specific statistical distribution, which is generated based on backbone conformational frequencies and excluded volume constraints (Jha, Colubri et al. 2005). In some of these cases, the statistical distribution is generated by using the frequencies of phi/psi dihedral angles taken from non-helix, non-turn, non-β strand residues within folded proteins. Statistical coil models have the advantage that they can generate many different conformations in a relatively short time, however, many of the initial conformations in the set may be energetically unfavorable as the method does not sample conformations according to a physically meaningful potential energy function (Cho, Nodet et al. 2009, Marsh and Forman-Kay 2009). Moreover, complex relations within the sequence, such as long-range contacts between residues that are distant in the primary sequence cannot be addressed simply by using these models. (The Flexible-Meccano algorithm provides an option to manually add prior knowledge modifications (Ozenne, Bauer et al. 2012)).

Once a structural library is generated, relative stabilities (or structural weights) need to be assigned to the various conformations. This can be done in a variety of ways; the most common

one is to select a subset of conformations that yields calculated observables that agree with experiment. It should be noted that selecting a subset of structures is equivalent to applying a particular weighting scheme to the structural library where structures that are not selected are assigned a weight of zero (Choy and Forman-Kay 2001, Chen, Campbell et al. 2007, Cho, Nodet et al. 2009, Marsh and Forman-Kay 2009, Fisher, Huang et al. 2010, Jensen, Salmon et al. 2010). The final set of structures and structural weights forms the structural ensemble. Some methods attempt to combine both the structure generation step and the assignments of structural weights in a single step. In one method the empirical potential energy function is modified to ensure that MD structures are sampled in a manner to yield ensemble averages that agree with experiment (Allison, Varnai et al. 2009, Esteban-Martin, Fenwick et al. 2009). This approach has yielded some promising results.

One inherent problem in generating a structural ensemble for IDPs is the degeneracy of the problem. In practice one tries to generate a structurally heterogeneous set of conformations for the structural library that captures, in some sense, the diversity of structures that are accessible to the IDP of interest. Weights are then assigned to yield calculated ensemble averages that agree with experiment. The ensemble generation procedure is therefore mathematically well defined. However, the number of experimental restraints that are available for a given IDP typically pales in comparison to the number of structures in the library. This leads to an inherently underdetermined problem; i.e., there are many different 'ensembles' (structures and weights) that yield calculated observables that are consistent with experiment.

Several methods were developed in the past in an attempt to handle the degeneracy problem. These can be divided into two main approaches. In one approach multiple different ensembles are generated for an IDP, all of which agree with experiments. Structural features that are shared among the different ensembles are then considered representative of the underlying real ensemble (Huang and Stultz 2008, Marsh and Forman-Kay 2009). In the second approach one first pre-defines a "prior" probability distribution, based for example on the potential energy of the conformations, and then an ensemble is generated such that it agrees with experiments as well as have maximal similarity with the prior. This is defined as the maximal entropy, or minimal information approach (Pitera and Chodera 2012, Boomsma, Ferkinghoff-Borg et al. 2014, Lane, Schwantes et al. 2014). While both methods have their merits, none of them provide quantitative estimates of the uncertainty in the final model.

In this thesis we employ a previously developed method, called Bayesian weighting (BW), for ensemble construction that quantifies one's uncertainty in the underlying ensemble. For a given structural library the method calculates a posterior probability distribution over all possible ways of weighting structures in the structural library. In practice, the BW ensemble consists of a structural library and the Bayes' estimate for the structure weights. A central feature of the approach is that the method calculates an "uncertainty parameter" that is associated with model correctness. An empirical study suggests that when the uncertainty parameter is 0, it is likely that the model is correct. By contrast, when the uncertainty parameter is equal to 1, there is little reason to have confidence that the resulting model is accurate. Nevertheless, even when one's uncertainty is high we can still provide confidence intervals for quantities that are calculated from the ensemble. In this sense, rigorous hypothesis testing can still be performed.

The Bayesian Weighting algorithm was used successfully on several different IDPs. In the first study a structural ensemble for K18 truncated form of tau protein (Fisher, Huang et al. 2010) was generated. In this study, chemical shifts, RDCs and Rg determined by SAXS were used to guide the generation of the posterior distribution. Since the uncertainty parameter was not equal to 0, confidence intervals were assigned to each of the predicted measurements that were made. In that study it was suggested that long-range contacts incorporate the PHF6* and PHF6, aggregation initiating hexapeptide motifs, within "hairpin like" formations, which may help protect the protein from aggregating (Fisher, Huang et al. 2010).

I.C α-Synuclein

α-Synuclein is a 140-residue IDP that is primarily expressed in presynaptic neurons throughout the central nervous system (Iwai, Masliah et al. 1995) and has been implicated as a causative agent in Parkinson's disease (PD), along with a number of other neurodegenerative diseases known collectively as synucleinopathies (Bellucci, Zaltieri et al. 2012). Over the last two decades there have been several key findings that suggest that α-synuclein plays a key role in both familial and sporadic PD pathogenesis (Venda, Cragg et al. 2010). Two genetic abnormalities were identified in early-onset familial form of PD: Duplication or triplication of the α-synuclein, *SNCA*, gene locus and three missense mutations (A53T, A30P and E46K) in the

SNCA gene. Recently two more mutations were discovered H50Q (Appel-Cresswell, Vilarino-Guell et al. 2013) and G51D (Lesage, Anheim et al. 2013). In addition, the pathological hallmark of PD is the existence of intraneuronal protein inclusions primarily found in the substantia nigra pars compacta known as Lewy bodies and Lewy Neurites (Lewy 1912). These aggregates are primarily composed of α-synuclein (Spillantini, Schmidt et al. 1997). Lastly, several animal models, where overexpressed wild type or mutant forms of α-synuclein were used, have neuronal inclusions, neurodegeneration and in some cases motor dysfunction akin to what is seen in PD (Whitworth 2011, Crabtree and Zhang 2012, Low and Aebischer 2012).

The α-synuclein sequence is divided into three regions: The N-terminal region (residues 1-60), the central region (residues 61-95) called the NAC (Non-Aβ component of Alzheimer's disease amyloid) and the C-terminal region (residues 96-140). The N-terminal region contains 4 (of the 7) 11-mer repeats with a KTKEGV motif. These repeats, reminiscent of apolipoproteins, can potentially form an amphipathic α-helix with a conserved acidic basic and hydrophobic arrangement (George, Jin et al. 1995). The NAC region contains the remaining repeats, it is highly hydrophobic and amyloidogenic. The C-terminal is acidic and is proline rich.

α-Synuclein has been referred to as a 'chameleon' due to its tendency to adopt different conformations under different conditions (Uversky 2003, Drescher, Huber et al. 2012). While α-synuclein is thought to predominantly exist as a disordered monomer in solution (Weinreb, Zhen et al. 1996, Fauvet, Kamdem et al. 2012), it can form secondary structure in different environments. The N-terminal region (residues 1-60) and non-amyloid-beta-component (NAC) region (residues 61-95) have shown a tendency to form amphipathic helices in association with micelles (Ulmer, Bax et al. 2005, Georgieva, Ramlall et al. 2008), supporting the notion that α-synuclein is involved in vesicle trafficking (Cooper, Gitler et al. 2006). The resulting helices appear to exist as two anti-parallel helices (Ulmer, Bax et al. 2005), or a single extended helix of variable length (Georgieva, Ramlall et al. 2008), depending on the size and composition of the vesicle with which it associates (Jao, Hegde et al. 2008). α-Synuclein fibrils have shown significant β-sheet secondary structure (Serpell, Berriman et al. 2000, Uversky, Li et al. 2001), as have certain protofibrillar oligomers, which are thought to lead to toxicity by increasing vesicle membrane permeabilization (Volles, Lee et al. 2001). The dual propensity of residues 1-95 highlights the fact that relatively ordered states, transient as they may be, also populate the

otherwise disordered conformational landscape of α-synuclein. Building structural ensemble for α-synuclein can potentially explain its chameleon nature.

Chapter II

Using Segmentation for Enhanced Sampling of Disordered Proteins

II.A Introduction

Intrinsically Disordered proteins (IDPs) are a class of proteins that, unlike archetypical folded proteins, are characterized by lacking a well-defined structure. These proteins therefore must be described by a heterogeneous set of rapidly interconverting conformations. IDPs have relatively flat energy landscapes that are comprised of multiple local energy minima that are separated by small barriers (Figure II.1). The shape of the energy landscape enables them to rapidly interconvert between different conformations (Uversky, Oldfield et al. 2008, Fisher and Stultz 2011).

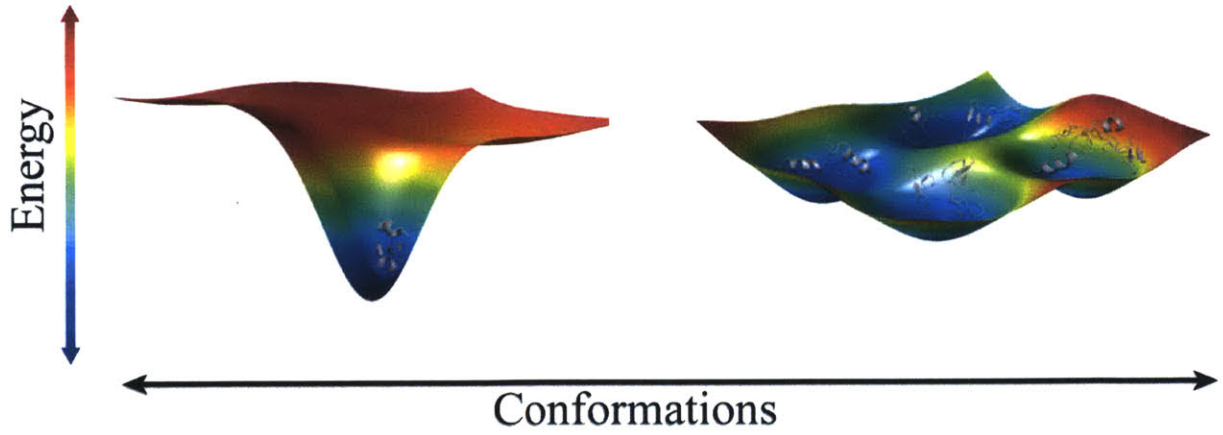

Figure II.1: Comparison of the energy landscape of a typical natively folded protein (on the left) and the one for a disordered protein (on the right).

Traditional experimental methods for studying folded proteins are often not applicable to IDPs. Experimental measurements on IDPs are typically made on time scales that are long with respect to the interconversion rates between distinct conformational states. Consequently experimentally measured quantities correspond to ensemble averages over a heterogeneous set of

structures. Using computational models designed to deconvolve these experimental observations is essential for developing a comprehensive understanding of these systems (Fisher and Stultz 2011).

One could generate a representative set of conformations that models accessible states of an IDP by directly sampling conformations from an empirical potential energy surface. However, exhaustive sampling for flexible polymers of just modest size is computationally prohibitive. For example, a freely jointed polypeptide chain of 140 residues can potentially sample approximately 3^{140} configurations, assuming that each monomer (amino-acid) can adopt one of three distinct states; e.g., helix, coil or strand. If it takes about one picosecond to explore each of these conformations computationally, it will take 10^{54} sec or more than 10^{47} years to explore all of them, thus a high-resolution detailed conformational description becomes computationally intractable. Therefore it is desirable to build "low resolution" models, consisting of a relatively small number of conformers, which capture the dominant thermodynamically accessible states of the protein. To this end, we define an ensemble as a finite set of energetically favorable conformations, denoted as $\{S_i\}$, and their relative stabilities $\{w_i\}$ (Fisher, Huang et al. 2010). Weight assignment is done in a manner that agrees with available experimental data (Figure II.2). One of the major challenges in building these models is generating a set of conformations that are sufficiently diverse to capture the range of conformations the protein can adopt.

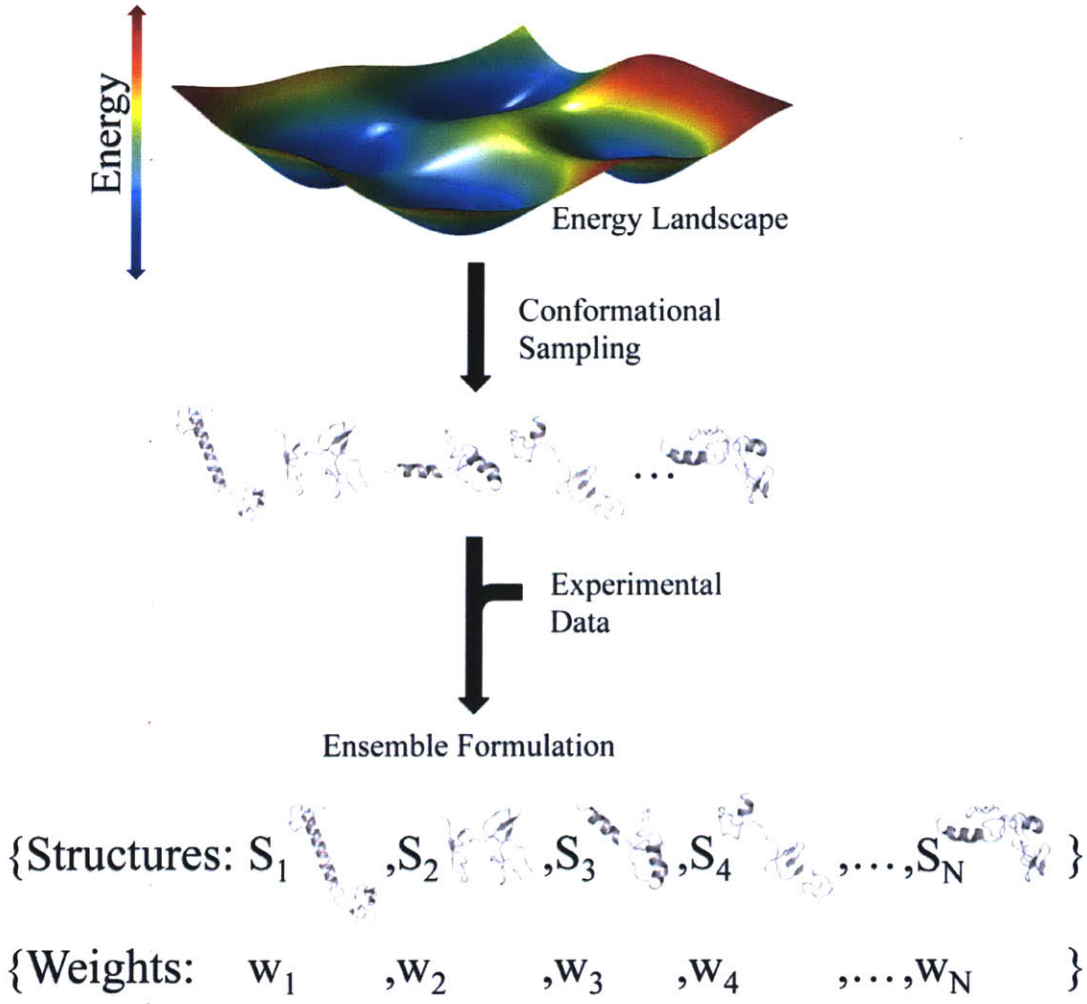

Figure II.2: Schematic showing our approach for building a conformational ensemble. The method entails sampling a set of energetically favorable conformations on a complex energy landscape, followed by assigning population weights to confomers.

There are several approaches for generating a heterogeneous set of conformations for a flexible protein. The first is a statistical coil based methodology - where backbone dihedral angles are sampled from the experimental dihedral propensity for a given amino acid. These propensities for the backbone dihedral angels are derived from regions within structured proteins that do not adopt well-defined secondary structure (Jha, Colubri et al. 2005, Ozenne, Bauer et al. 2012). In this sense, the potential energy surface is derived from empirically calculated statistics from folded proteins – an approach that is not necessarily appropriate to describe an IDP. In addition this statistical "random coil" description on its own does not allow for long-range contacts, which were shown to be important within α-synuclein for example (Bernado,

Bertoncini et al. 2005). Other approaches utilize physical, but empirical, potential energy functions that model interactions between atoms in the protein as well as protein-solvent interactions. Sampling on this potential energy surface is typically done with either Monte Carlo or Molecular Dynamics (MD) simulations (Karplus and McCammon 2002, Rauscher and Pomes 2010). In this work we chose a technique, named Replica Exchange Molecular Dynamics (REMD) simulations, that allows for enhanced sampling of the energy landscape (Sugita and Okamoto 1999), in a reasonable amount of CPU time. We apply these methods to create a structural ensemble for α-synuclein, a 140 amino acid long IDP that has been suggested to play a role in the pathogenesis of Parkinson's disease.

II.B Results and Discussion

II.B.1 Generating a Library of Energetically Favorable Conformations

To create a structural library we need to generate a set of energetically favorable, yet structurally diverse, conformations. In this endeavor it is important to first decide on an unambiguous measure that quantifies the diversity in our structural library. One can quantify the structural diversity of a set of conformations using many different metrics. In this work we chose the radius of gyration, a measure of structure compactness, as a metric to quantify diversity. While we strive to achieve diversity in the range of radii of gyration in the structural library, we demonstrate that with this criterion we also obtained structures of diverse secondary structure content.

The most straightforward way to generate a set of energetically favorable conformations is to run MD simulations with the full protein sequence, thereby allowing the system to sample states that are accessible at the chosen temperature. As mentioned above, while this approach is fruitful for folded protein it is impractical for IDPs. We therefore used a common technique for enhanced sampling (REMD) (Sugita and Okamoto 1999). In this method several copies, or replicas, of the protein run in parallel at different temperatures. After a certain number of MD steps, conformations from different replicas are allowed to swap according to the usual Metropolis criterion, i.e. a transition is made with probability $max\{1, exp[-\Delta E/k_B(-\Delta T)]\}$, where ΔE and ΔT are the energy difference and temperature difference, respectively, between

two replicas. Simulations were run with an implicit solvent model. Using implicit solvent model reduces the degrees of freedom in the system, relative to explicit solvent models, and therefore yields faster calculations. In this work we used the EEF1 implicit solvent model (Lazaridis and Karplus 1999) and the CHARMM force field (Brooks, Bruccoleri et al. 1983) for our replica exchange simulations. In a study comparing different implicit solvent models to the TIP3P explicit solvent model (Neria, Fischer et al. 1996, Lazaridis and Karplus 1999) using a six residue amyloidogenic peptide, it was found that EEF1 implicit solvent model generally reproduces the set of energy minima sampled using a TIP3P model of explicit solvent (Lazaridis and Karplus 1999, Steinbach 2004, Huang and Stultz 2007, Strodel and Wales 2008).

We ran a 10ns REMD simulation using the full α-synuclein sequence (see methods for details). The total CPU time was approximately 36 hours for this simulation. Structures from the room temperature replica were collected every picosecond from the last 5ns to generate a library consisting of 5000 structures.

As can be seen in Figure II.3 these simulations yielded radii of gyration in a narrow range of values; i.e., structures arising from the simulation are almost as compact as what one would expect from a folded protein having the same amino-acid length (Figure II.3A) (Gast, Damaschun et al. 1995). To further assess the extent of structural heterogeneity we computed the root-mean-square deviations (RMSD) between all pairs of structures in the library. There a non-negligible fraction of the structures are very similar; e.g., approximately 14% of the pairs have an RMSD less than 5Å (Figure II.3B).

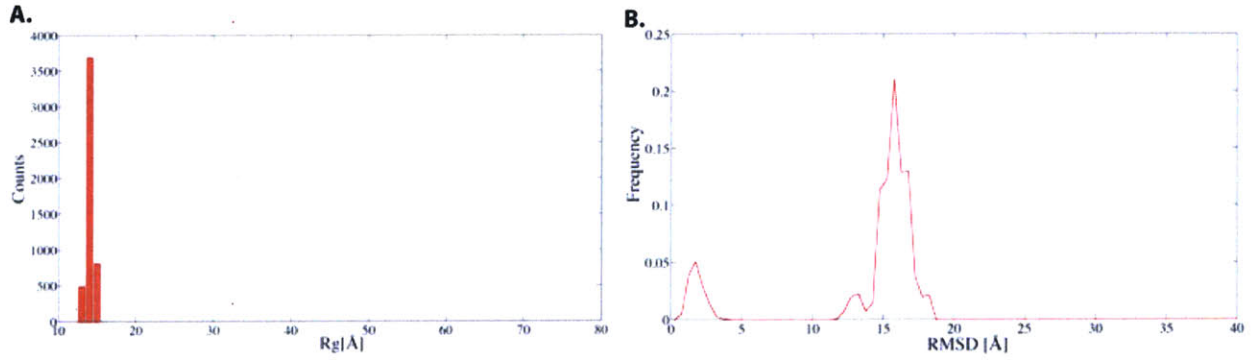

Figure II.3: Full-length Protein Structural Library Diversity –A. Radius of gyration calculated from the 5000 structures that make the structural library of the full-length protein REMD simulations. B. Pair-wise RMSD for each unique pairing within the structural library.

These data suggest that simulating the full-length protein with REMD yields a structural ensemble that has a limited amount of heterogeneity. Moreover, generating these data required more than 36 hours of simulation time. We therefore examined an alternate approach structure generation. The method divides the protein into smaller segments and then exhaustively samples the conformational space of each segment. Structures for the full-length protein are then constructed by piecing together conformations of the individual segments. By running molecular dynamics over segments compared to the full-length protein, we are reducing the size of the system, thereby shortening the simulation time. Since segments simulations can be done in parallel we can sample the full-length conformational space in a fraction of the CPU time. The use of protein fragments to deduce the structure of the full-length protein has shown success both experimentally (Marqusee, Robbins et al. 1989, Shin, Merutka et al. 1993, Waltho, Feher et al. 1993, Blanco, Rivas et al. 1994, Zerella, Evans et al. 1999, Eliezer, Chung et al. 2000) and computationally (Bystroff and Garde 2003, Ho and Dill 2006, Voelz, Shell et al. 2009) for folded proteins.

To this end we decided to divide the sequence of α-synuclein into eight residue long overlapping segments resulting in 28 segments in total (Figure II.4) (the C-terminal segment was five residues long). The size of the segments (8 residues) was chosen based on the experimental average persistence length of a denatured polypeptide (Damaschun, Damaschun et al. 1993, Schwalbe, Fiebig et al. 1997). In addition, it was shown that for folded proteins, in some cases, one can independently sample 8 long peptide fragments and reproduce the native structure of these fragments within the context of the whole protein (Ho and Dill 2006). This suggests secondary structure elements can be captured by short sequences when non-local interactions can be neglected. Each segment had three overlapping residues with adjacent segments as can be seen in Figure II.4.

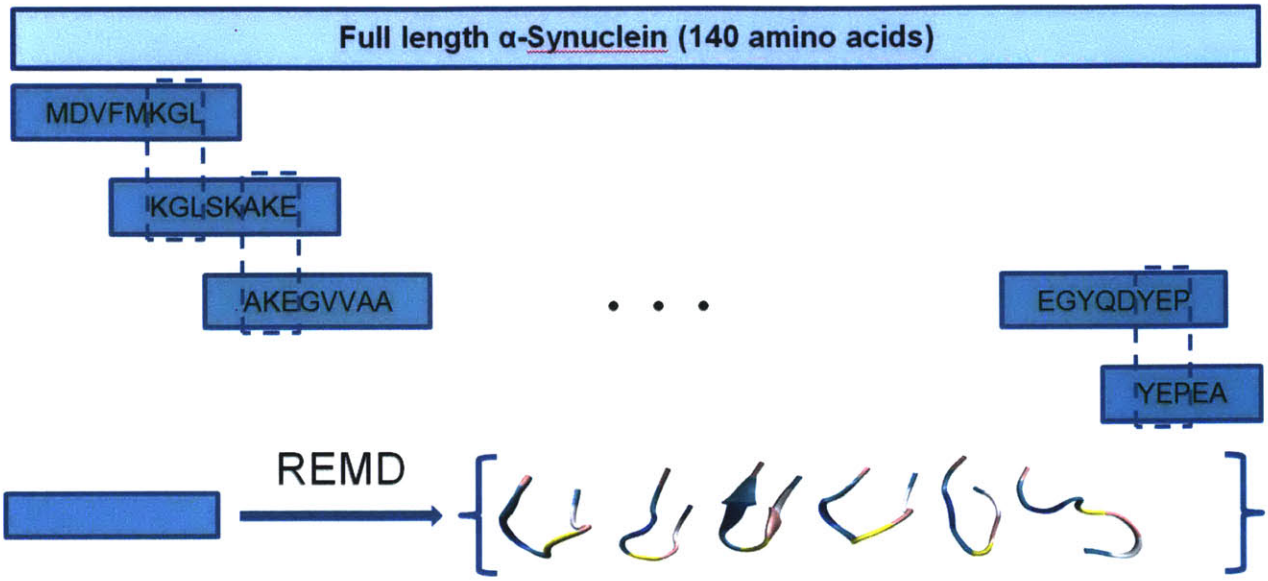

Figure II.4: A schematic describing the process of segmenting the protein. The full-length sequence is divided into 8 aa-long segments where each neighboring segments overlap with 3 residues; the last segment is only 5 residues long. Each of these segments undergoes sampling using REMD

Full-length α-synuclein conformations were generated by piecing together segments one at a time, starting with the N-terminal segment. Each segment was clustered using the three overlapping residues at its ends. The segment to be added to the growing polypeptide chains was chosen from the cluster that had the most structural similarity in the overlapping region. A similar protocol was used to describe K18, an intrinsically disordered protein of comparable size, 130 amino acids long (Fisher, Huang et al. 2010).

The first residue coordinates of the overlapping segments were taken from the C-terminal region of the one segment and the two others from the N-terminal region of the adjoining segment. At the end of the procedure, the full length structure was subjected to 1000 steps of steepest descent minimization followed by 10,000 steps of adopted basis Newton–Raphson minimization with the EEF1 implicit solvent model (Lazaridis and Karplus 1999) to relieve any bad contacts in the molecule. Moreover, minimization of the combined full length conformations helps to re-introduce, in part, long range interactions that are ignored by segmenting the protein and independently sampling each segment. The procedure is depicted in Figure II.5. Only structures with a negative energy were chosen for the structural library. This resulted in a structural library of approximately 80,000 structures.

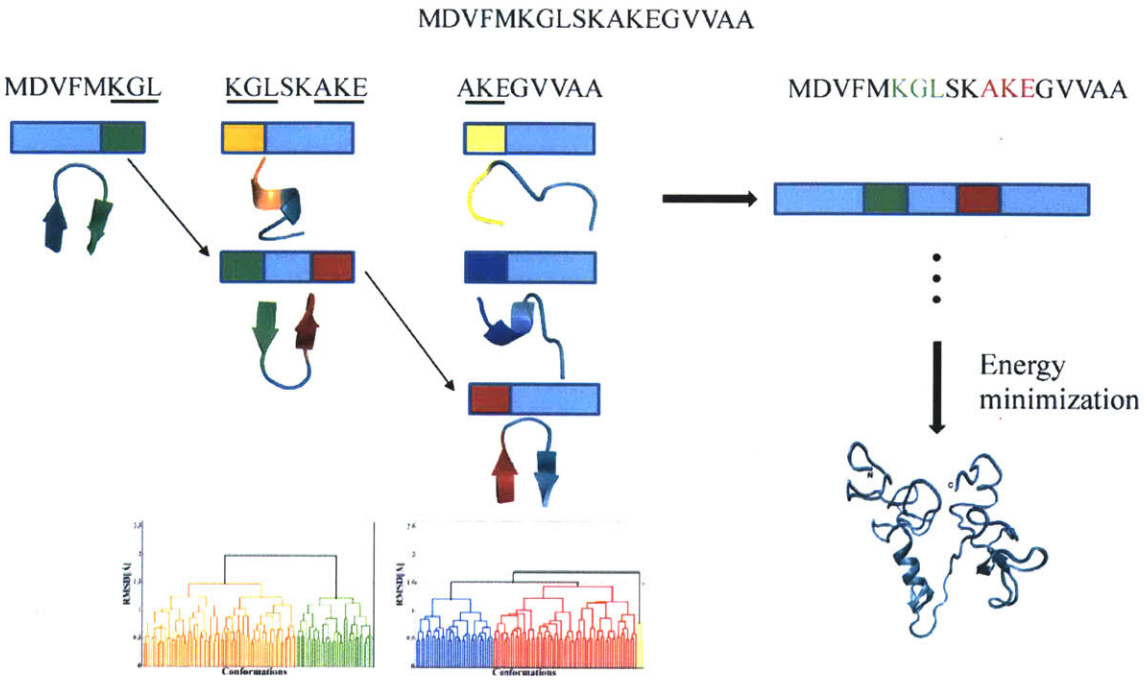

Figure II.5: Combining the segments together was performed to increase the conformational similarity of the overlapping residues. This figure shows the dendograms generated by clustering using the pairwise RMSD of the first three residues of each segments. For example, there are only two clusters in the second segment, the green and the orange, from each of these clusters we randomly choose a representative. The first segment is the combined with the representative that has the highest similarity in the overlapping residues. This process is repeated until a full-length α-synuclein conformation is built. The structure composed of the combined segments is then subjected to energy minimization.

The total CPU time for an 8 aa-long segment took 8.5 hours (compared to the 36 of the full-length protein simulation). Since these simulations are run independently, and in parallel, one can effectively simulate the entire protein in under 9 hours. However additional time is required for the minimization procedure after the protein was pieced together. For the 80,000 structures minimization added additional ~2.5 hours using an architecture that allows us to run all the segments in parallel, i.e. 28 nodes with 16 virtual cores (1 virtual core for each replica). In sum, simulating the full-length approach produced 5000 conformations within 36 hours while the segmentation approach yielded 80,000 conformations within ~11 hours.

We calculated how the segment-based approach compares with the full-length approach in terms of heterogeneity. Due to the size difference of the structural libraries, i.e., 80,000 structures compared to 5,000 structures. We find that using the segment-based approach achieves more diversity with respect to the radius of gyration (Figure II.6A). The pair-wise

RMSDs suggest that the structures are more diverse; i.e., the peak at <5 angstroms that was present in the full-length data is not present in data arising from the segment-based approach.

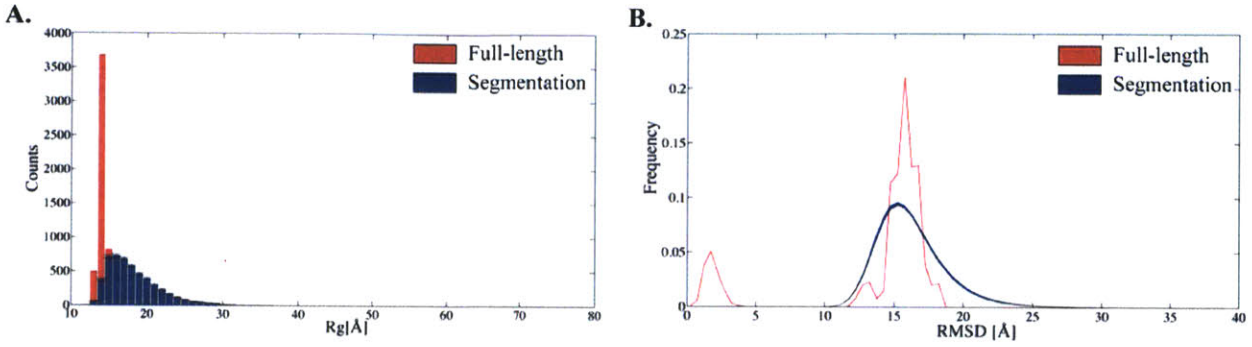

Figure II.6: A. Radius of gyration (R_g) calculated from the REMD of the full-length vs the segment-based approach. Since the segment based approach yielded 80,000 structure and the full-length approach yielded only 5000 structures, we randomly selected a sample of 5,000 structures from the segment-based approach library and calculated the R_g and the pair-wise RMSD distribution for that sample. The random sampling process was repeated a 100 times and each bin represents the mean $\pm\sigma$. B. Pair-wise RMSD distribution for a simulation of the full-length protein (red) compared to the segment-based simulation (blue).

II.B.2 Evaluating the Structural Library

Using a segment based approach generates a structural library that is more diverse than that generated from simulations with the full-length protein. However, it is not clear from these data alone whether the range of radii of gyration is sufficient for our purposes. To estimate the range of radii of gyration that should be sampled we rely on previous observations made by Flory and Fiske (Flory 1953, Flory and Fisk 1966). To estimate the upper bound associated with the expected gyration, we rely on the calculated distribution of the R_g for a freely jointed chain that has a mean value of a self-avoiding polymer chain – a value obtained from the Flory power law (Flory 1953, Flory and Fisk 1966) (see methods section). The lower bound is calculated using an empirical formula for compact globular proteins (Gast, Damaschun et al. 1995) (see methods section). The resulting radius of gyration range is between 16Å and 71Å. As can be seen in Figure II.7, the R_g histogram over 80,000 structures did not cover the desired R_g range.

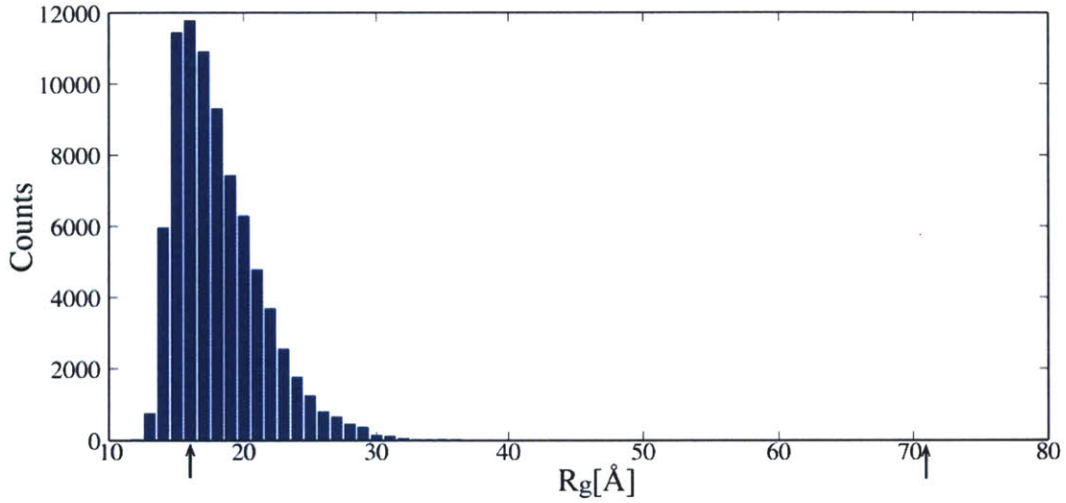

Figure II.7: Histogram of the radius of gyration for the 80,000 structures generated using the segment based approach and energy minimization of the combined structures. The arrows denote the predefined R_g range required for structural diversity.

In order to cover a wider range of radii of gyration, these 80,000 structures were subjected to additional energy minimizations using a R_g restraint. R_g restraints varied from 27 to 75 Å. This procedure resulted in approximately 300,000 structures. Figure II.8 shows the distribution of the radius of gyration of this expanded structural library.

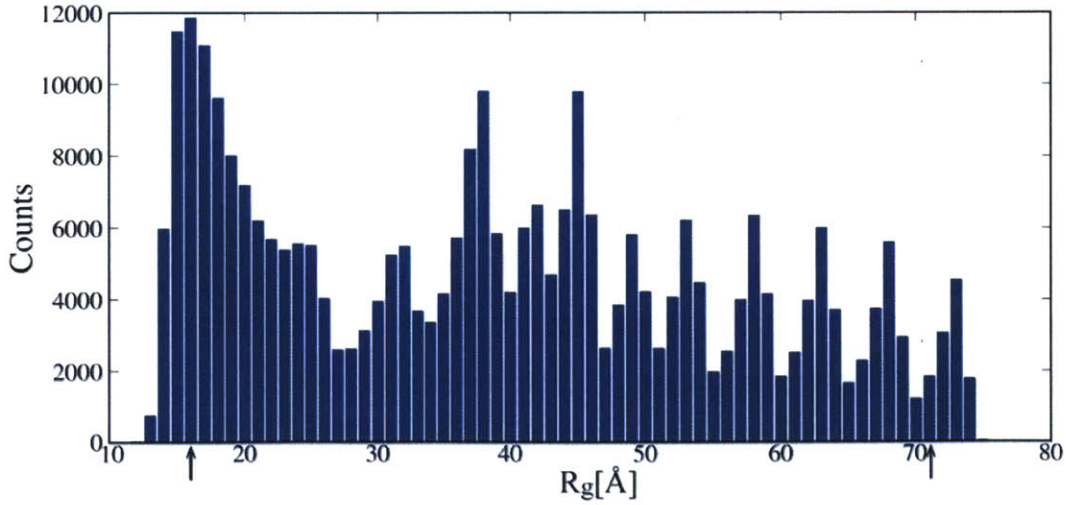

Figure II.8: Histogram of the radius of gyration for the 300,000 structures generated using the segment based approach and energy minimization of the combined structures. The arrows denote the predefined R_g range required for structural diversity.

In addition to spanning a wide range of radii of gyration, structures in our structural library also sample varying amounts of secondary structure content (Figure II.9).

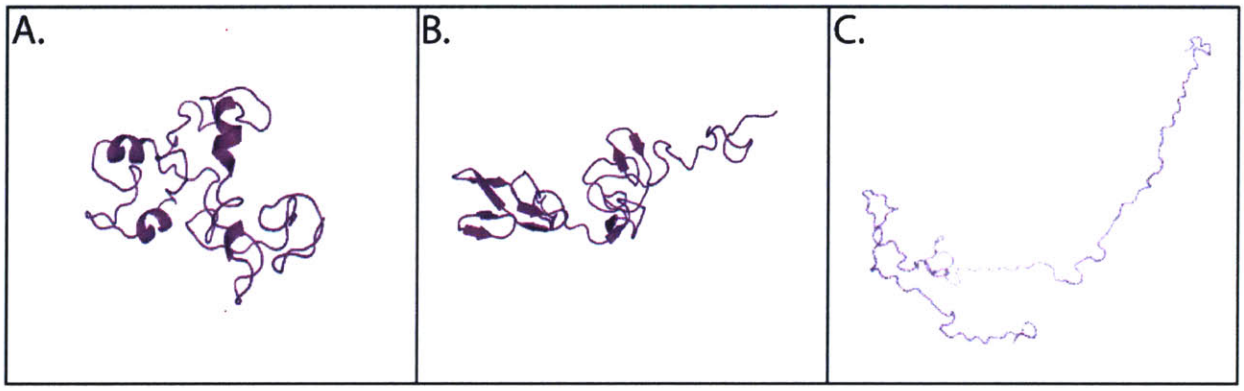

Figure II.9: Secondary structure diversity: examples of structures that have the most helical content (A.), the most strand content (B.) and an example of an unstructured conformation (C.).

While the method we described provides an efficient way to obtain a diverse set of structures, the scheme also carries some disadvantages. For example, the rapid method for piecing together different sequence fragments also led to structures that contained "knots" (Virnau, Mirny et al. 2006) (see example in Figure II.10).

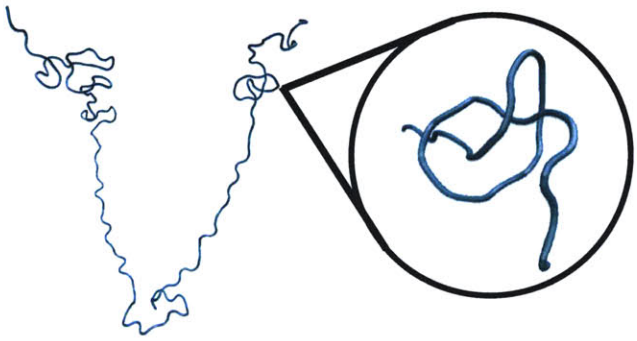

Figure II.10: A close up look at a structure containing a trefoil knot (3_1).

While knots have been observed in a very small number of folded proteins (Dzubiella 2009), it is unlikely that they would be sampled by disordered proteins. Indeed, it has been shown that knots in folded structures tend to persist even under denaturing conditions (King, Jacobitz et al. 2010). Thus knots present an irreversible process, one that will likely hinder the

fast interconversion of conformers that is characteristic of IDPs. We therefore reasoned that knots should not be present in IDPs and used the program KNOTS to automatically identify structures containing knots and removed them from the library (Kolesov, Virnau et al. 2007). The speed in which knots can be identified made it more efficient to generate structures using a fast segment-based method followed by removal of knotted structures. After removing these structures we were left with a structural library of ~100,000 structures. As we see in Figure II.11, the resulting library spans the R_g range we predefined to indicate diversity.

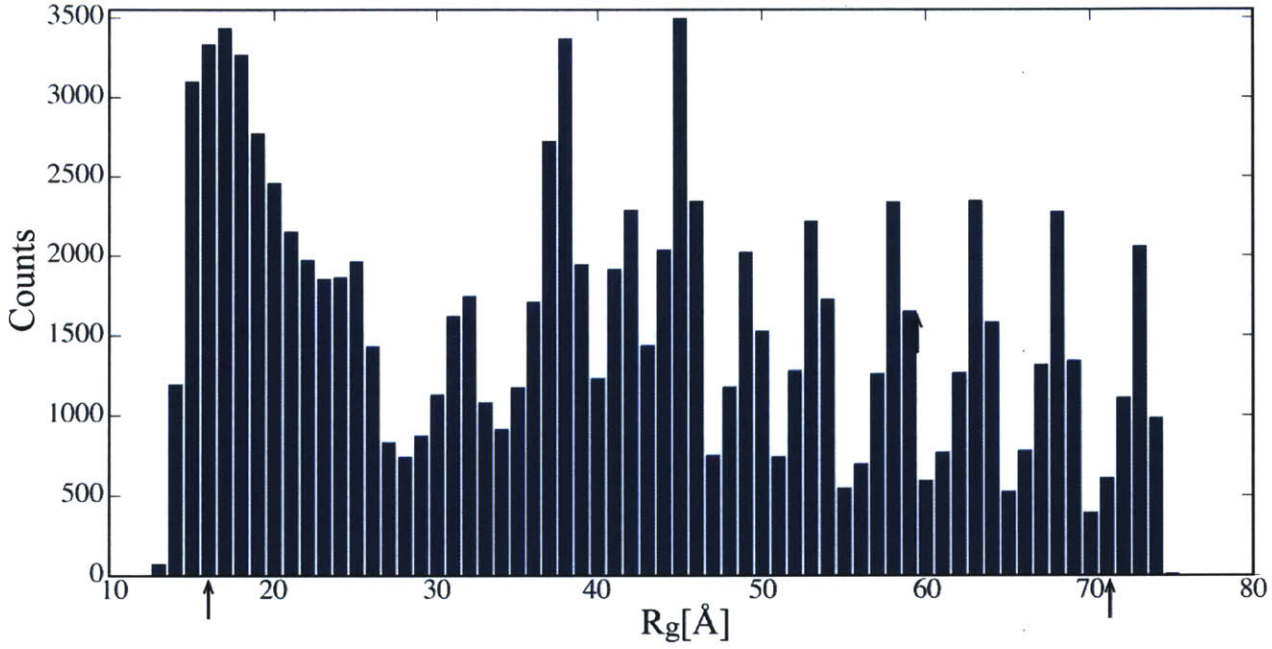

Figure II.11: Histogram of the radius of gyration for the structural library that resulted from energy-minimized combined structures in addition to R_g-restrained minimization and removal of knotted structures. The arrows denote the predefined R_g range required for structural diversity.

The structural library was reduced in size to 299 structures (see Figure II.12) that largely captures the structural diversity present in the larger library that contained 100,000 structures. This was achieved using a previously described pruning algorithm (Fisher, Huang et al. 2010). (A similar procedure was done to build a model for the K18 tau segment of comparable size (130 residues)) (Fisher, Huang et al. 2010). Reducing the size of the library facilitates the estimation of structural weights – a process needed to arrive at the final ensemble.

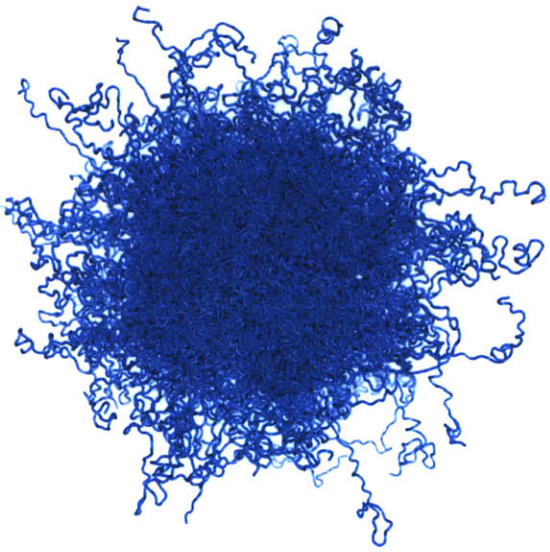

Figure II.12: An alignment of all 299 structures that were used to construct the α-synuclein ensemble.

Once we have obtained a diverse structural library, we were left with the task of determining the relative stabilities of the structures within. These relative stabilities are determined with the aid of experimental data. However, since the number of experimental constraints pales in comparison to the number of degrees of freedom we are left with an underdetermined task for which many combinations of relative stabilities and structures can lead to ensembles that agree with the experimental experiment. In the next chapter we will discuss how one is able to use a Bayesian methodology that produces a distribution, over all the possible sets of weights for a given set of structures and experimental observables.

II.C Methods

II.C.1 Radius of gyration calculations

According to Flory the R_g of a random coil follows a simple scaling law:

$$\langle R_g \rangle = R_0 N^{\upsilon}$$

Where N is the number of monomers in the polymer chain (in our case 140). The values for the constant R_0 and the scaling factor υ were obtained from an extensive SAXS study of 28

proteins under strong denaturing conditions, where the best-fit values for the power law are: $R_0 = 1.927^{+0.271}_{-0.238}$Å and $\upsilon = 0.598 \pm 0.028$ and the bounds represent 95% confidence intervals (Kohn, Millett et al. 2004, Kohn, Millett et al. 2005). To extend the range for the probability distribution we chose to use the higher bounds for the formula of the mean radius of gyration (i.e. $R_0 = 2.198$Å and $\upsilon = 0.626$). Therefore the calculated average radius of gyration is $\langle R_g \rangle = 48.5$Å.

The empirical distribution of the R_g is (Flory and Fisk 1966):

$$P(R_g) = A(R_g{}^2)^3 exp\left(-\frac{7}{2}\left(\frac{R_g{}^2}{\langle R_g{}^2 \rangle}\right)\right)$$

Where A is a normalization constant. It can be shown that under this distribution $\langle R_g{}^2 \rangle \cong 1.08\langle R_g \rangle^2$ (see Appendix for detailed derivation) and therefore the distribution formula is fully defined and we can sample from it and calculate the 95th percentile as the upper bound for the R_g. To calculate the lower bound we used an empirical formula for the R_g of globular proteins $\langle R_g \rangle = 2.9N^{1/3}$ where N is again the number of residues (Gast, Damaschun et al. 1995).

II.C.2 Replica Exchange Molecular Dynamics simulations

The following protocol was used for both the full-length protein simulations and the segments simulations.

Molecular dynamics simulations were performed using the CHARMM force field (Brooks, Bruccoleri et al. 1983) with EEF1 implicit solvent (Lazaridis and Karplus 1999). A total of 16 replicas, each at a different temperature, were used. Temperatures were spaced exponentially in the range 280–700 K. Simulations were run for 10 ns, and structures were collected from the last 5 ns of the 298 K heat bath, allowing 5 ns of equilibration period. A total of 5000 conformations per simulation were collected. Each conformation was subjected to a short minimization procedure of 500 steps of adopted basis Newton–Raphson, again with the EEF1 implicit solvent model (Lazaridis and Karplus 1999).

Chapter III

Explaining the Structural Plasticity of α-Synuclein

III.A Abstract

Given that α-synuclein has been implicated in the pathogenesis of several neurodegenerative disorders, deciphering the structure of this protein is of particular importance. While monomeric α-synuclein is disordered in solution, it can form aggregates rich in cross-β structure; relatively long helical segments when bound to micelles or lipid vesicles; and a relatively ordered helical tetramer within the native cell environment. To understand the physical basis underlying this structural plasticity, we generated an ensemble for monomeric α-synuclein using a Bayesian formalism that combines data from NMR chemical shifts, RDCs and SAXS with molecular simulations. An analysis of the resulting ensemble suggests that a non-negligible fraction of the ensemble (0.08, 95% confidence interval 0.03-0.12) places the minimal toxic aggregation-prone segment in α-synuclein, NAC(8-18), in a solvent exposed and extended conformation that can form cross-β structure. Our data also suggest that a sizeable fraction of structures in the ensemble (0.14, 95% confidence interval 0.04-0.23) contains long range contacts between the N- and C-termini. Moreover, a significant fraction of structures that contain these long range contacts also place the NAC(8-18) segment in a solvent exposed orientation – a finding in contrast to the theory that such long range contacts help to prevent aggregation. Lastly, our data suggest that α-synuclein samples structures with amphipathic helices that can self-associate via hydrophobic contacts to form tetrameric structures. Overall, these observations represent a comprehensive view of the unfolded ensemble of monomeric α-synuclein and explain how different conformations can arise from the monomeric protein.

This chapter was published in similar form in: Ullman, O., Fisher, C.K., and Stultz, C.M. Explaining the Structural Plasticity of α-Synuclein. *J Am Chem Soc* **2011**, *133*, 19536-19546

III.B Introduction

α-Synuclein is a 140 amino acid protein that has been implicated in several neurodegenerative diseases, often referred to as synucleopathies, such as Parkinson's disease (PD), Dementia with Lewy bodies (DLB) and Multiple System Atrophy (MSA) (Spillantini and Goedert 2000, Galvin, Lee et al. 2001, Goedert 2001). PD, in particular, is neuropathologically characterized by α-synuclein aggregates and the loss of dopaminergic neurons within the substantia nigra (Forno 1996, Spillantini, Crowther et al. 1998). While a number of theories have been advanced to explain how α-synuclein self-association is related to neuronal dysfunction, the precise relationship between α-synuclein aggregation and cell death remains unclear (Maries, Dass et al. 2003). Consequently, understanding the structural basis of α-synuclein self-association is of particular importance.

Although monomeric α-synuclein is intrinsically disordered in aqueous solution and is therefore considered an intrinsically disordered protein (IDP), it cannot be simply described as a random coil (Weinreb, Zhen et al. 1996, Eliezer, Kutluay et al. 2001, Uversky 2003). For example, the average radius of gyration of a random coil that is 140 amino acids long is larger than the measured average radius of gyration for α-synuclein obtained via small angle x-ray scattering (SAXS) experiments (Li, Uversky et al. 2001). This suggests that α-synuclein is, on average, more compact than the classic random coil.

In addition, α-synuclein can form ordered structures under different experimental conditions. The amino acid sequence of α-synuclein contains 11-residue imperfect repeats that are distributed among the highly basic N-terminal region of the protein (residues 1-60), and the hydrophobic NAC region (Non-Aβ Component of α-synuclein, residues 61-95). These repeats were proposed to form amphipathic α-helices capable of interacting with different types of lipid structures (George, Jin et al. 1995, Davidson, Jonas et al. 1998). It was found that when α-synuclein is bound to micelles, two helices can form (Bussell and Eliezer 2003, Chandra, Chen et al. 2003). The first helix encompasses residues 3 to 37 and is therefore contained within the N-terminal region and the second helix is formed between residues 45 and 92 – a region that begins in the N-terminal region and extends into the NAC region. The two helices are aligned antiparallel to one another (Ulmer, Bax et al. 2005, Borbat, Ramlall et al. 2006). Other studies suggest that α-synuclein can also form a continuous helix that begins in the N terminal and

continues through to the NAC region and that the precise form of the helical segment depends on the precise experimental conditions (Georgieva, Ramlall et al. 2008, Jao, Hegde et al. 2008, Ferreon, Gambin et al. 2009, Trexler and Rhoades 2009). For example, in a recent study it was found, using pulsed dipolar ESR spectroscopy, that depending on the relative protein-to-detergent concentrations, α-synuclein can adopt either a single extended helix form or the broken helix form, similar to the one previously described (Georgieva, Ramlall et al. 2010). Moreover, recent data further suggests that under physiologic conditions, α-synuclein exists in a tetrameric form that has considerable helical content (Bartels, Choi et al. 2011).

By contrast, α-synuclein aggregation is characterized by an increase in β-sheet content. Atomic force microscopy and Raman spectroscopy demonstrated that soluble α-synuclein oligomers have reduced α-helical content relative to protofilaments, and that the β sheet content is relatively increased in protofilaments and filaments (Apetri, Maiti et al. 2006). Fiber diffraction of α-synuclein fibrils further demonstrated the presence of cross-β structure which is characteristic of amyloid fibrils (Serpell, Berriman et al. 2000). Consequently, the available experimental evidence suggests that α-synuclein can adopt helical structures or extended structures depending on the binding partner and experimental conditions.

A number of studies have constructed α-synuclein ensembles, using a combination of computational methods and experiments, to better understand the nature of the unfolded state (Bernado, Bertoncini et al. 2005, Dedmon, Lindorff-Larsen et al. 2005, Allison, Varnai et al. 2009, Koo, Choi et al. 2009, Wu, Weinstock et al. 2009). Some of these studies combine data obtained from NMR, PRE, and conformational sampling to construct an appropriate ensemble (Bernado, Bertoncini et al. 2005, Dedmon, Lindorff-Larsen et al. 2005, Allison, Varnai et al. 2009, Wu, Weinstock et al. 2009). While these studies have provided insights into the accessible states of α-synuclein in solution, there are still many unanswered questions regarding the unfolded state of this protein, including the precise role of secondary structure in the unfolded ensemble and the presence of long range contacts, in particular. In addition, the recent observation that α-synuclein can also form ordered helical tetramers in the native cell environment has not been addressed in the previous studies.

In this work we use a recently developed Bayesian Weighting (BW) algorithm to construct an ensemble for wild-type (WT) α-synuclein (Fisher, Huang et al. 2010). Data from NMR chemical shifts (Rao, Kim et al. 2009), RDCs (Bertoncini, Fernandez et al. 2005) and

SAXS (Binolfi, Rasia et al. 2006) experiments are used to guide the construction of the ensemble. An analysis of the ensemble: 1) helps to clarify the role of secondary structure propensity and the different binding characteristics of α-synuclein; 2) identifies potential aggregation prone structures within the ensemble; 3) clarifies the relationship between long range contacts and aggregation propensity; and 4) provides insights into how the disordered monomeric protein can form tetrameric helical structures.

III.C Results and Discussion

III.C.1 Construction of an α-Synuclein Ensemble

We began by generating a relatively large structural library of energetically favorable conformations and then used a Bayesian weighting (BW) algorithm (Fisher, Huang et al. 2010) to assign weights (or relative stabilities) for each conformer in the library. Hence an 'ensemble' is defined as a set of structures, $\{S_i\}$ and a corresponding set of weights $\vec{w} = \{w_i\}$ where w_i is the weight (or probability) of structure S_i and $\sum_i w_i = 1$.

For a given structural library there are many possible ways to weight the different structures within the structural library, and each possible weighting scheme represents a different ensemble. The BW method assigns a probability to every possible weighting scheme, and hence every possible ensemble, that can be constructed from the structural library. Parenthetically we note that since some of the w_i can be 0, finding a correct weighting scheme also enables us to eliminate exclude structures from the structural library if they consistently lead to ensembles that are inconsistent with the experimental data.

The probability of a given ensemble is calculated using methods from Bayesian statistics as described in our previous work (Fisher, Huang et al. 2010) and as reviewed in the Methods. Overall the probability of an ensemble is related to the agreement between the data predicted by the ensemble and the experimental data. In the Bayesian formalism we compute a probability distribution (which we refer to as the posterior density) over all possible ensembles, and this distribution is used to make statements about the conformational properties of α-synuclein. Since the posterior density is a multi-dimensional function, we summarize its properties in two ways. First, we calculate the average weight of each structure in the structural library using the

posterior density function. The ensemble consisting of the structures $\{S_i\}$ and these average weights, $\vec{w}^B = \{w_i^B\}$ is called the Bayes ensemble; i.e., it is the Bayesian analogue of a 'best fit' ensemble. Of course, the average of a distribution may not be very informative if the standard deviation, a measure of uncertainty, is large. To reflect this, we use the distribution over ensembles (the posterior density) to calculate confidence intervals for conformational characteristics of α-synuclein as a way of quantifying statistical uncertainty. Note that the confidence intervals do not refer to a specific ensemble, but rather to the distribution over all possible ensembles that could be constructed from the structural library.

An advantage of the BW formalism is that it provides a built in estimate of the uncertainty in the Bayes ensemble. Since agreement with experiment alone does not ensure that an ensemble is correct, such quantitative measures of uncertainty are important (Fisher, Huang et al. 2010). A further advantage of the method is that even when the uncertainty in the Bayes ensemble is relatively large, we can calculate error bars to quantify the uncertainty in any observable quantity that is calculated from the ensemble.

First, we constructed a structural library of 100,000 energetically stable structures by breaking the protein into overlapping 8-residue segments and exhaustively sampling the conformational space of each segment using Replica Exchange Molecular Dynamics (REMD) (Sugita and Okamoto 1999). Segments were then joined to form a structure of the full 140 residue protein. To reduce the number of conformations to a more manageable size, the structural library was pruned using a coarse clustering method to generate a set of 299 structures that largely preserves the structural heterogeneity that was present in the original structural library (Figure III.1).

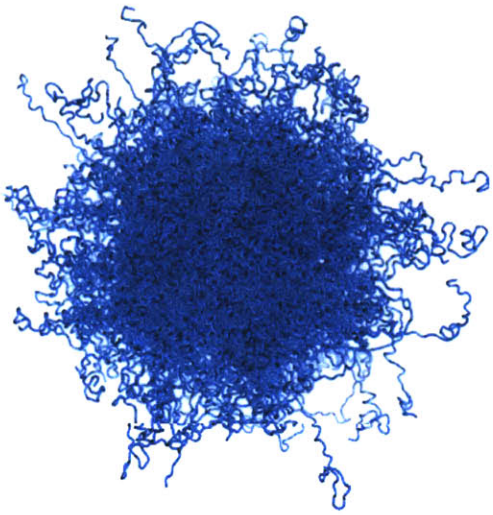

Figure III.1: An alignment of all structures within the α-synuclein ensemble.

Application of the BW algorithm to obtain the Bayes weight for each structure yielded a Bayes ensemble that agrees with measured NMR chemical shifts (Rao, Kim et al. 2009) (Figure III.2A) and RDCs (Figure III.2B) (Bertoncini, Fernandez et al. 2005) as well as SAXS derived radius of gyration (Binolfi, Rasia et al. 2006) (ensemble average value 41±1Å vs. experimentally determined value of 40±2Å). These data demonstrate that the BW algorithm accomplishes its goal of generating ensembles that agree with the input experimental data.

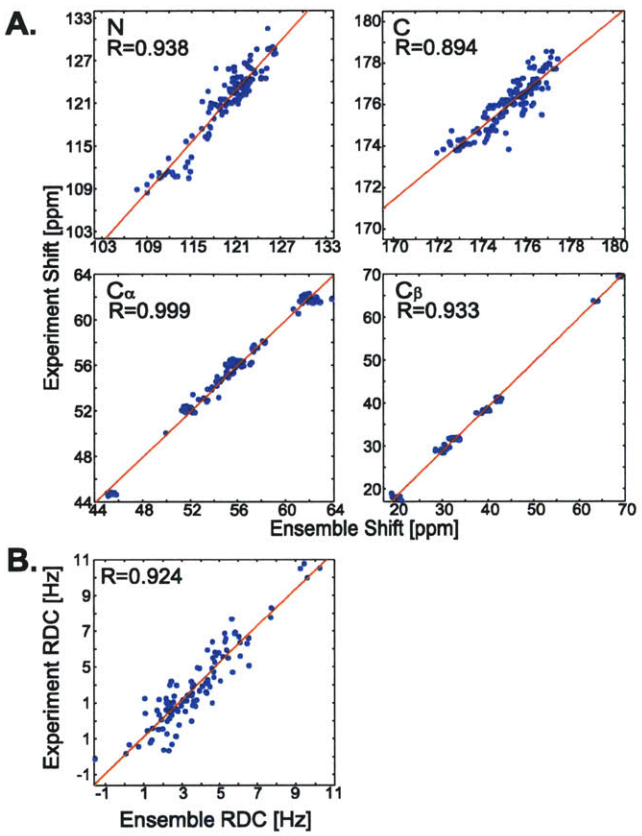

Figure III.2: Ensemble Agreement with Experimental Data. Comparison of experimental results with the corresponding calculated (A) chemical shifts and (B) RDCs. Correlation coefficients are explicitly shown. Calculated Root Mean Square Error (RMSE) for the Chemical shifts was found to be within accuracy provided by SHIFTX (Neal, Nip et al. 2003)

As discussed above, the BW method provides a built in metric, called the the uncertainty parameter, that quantifies our uncertainty in the Bayes ensemble, and is analogous to the standard deviation of a Gaussian distribution (Fisher, Huang et al. 2010). If it is likely that the Bayes ensemble is correct, the uncertainty parameter approaches 0. Conversely, if it is unlikely that the Bayes ensemble is correct, then the uncertainty parameter approaches 1. In other words, as the uncertainty parameter approaches 1, we cannot say with any certainty that the constructed ensemble is correct. In the present case, the uncertainty parameter is 0.4. In this scenario, we can further quantify our uncertainty by computing confidence intervals for specific conformational characteristics.

An analysis of the Bayes ensemble provides additional information about the relative distribution of different conformer sizes that are accessible to the protein. As shown in Figure III.3, the ensemble itself contains structures with radii of gyration that range from approximately

20Å to 60Å. To put these values into perspective we note that the average radius of gyration for a globular folded protein containing 140aa is approximately 15Å, while the average radius of gyration for a random coil with the same amino acid length is approximately 52Å (Uversky, Li et al. 2001). The fraction of the ensemble with a radius of gyration near 20Å is 0.09 (95% confidence interval 0.05-0.19) while the fraction that has a radius of gyration greater than that would be expected based on the random coil calculation is 0.17 (95% confidence interval 0.13-0.22). This suggests that α-synuclein samples structures that are nearly as compact as a globular protein of the same size in addition to structures that are more extended than that of the average random coil value.

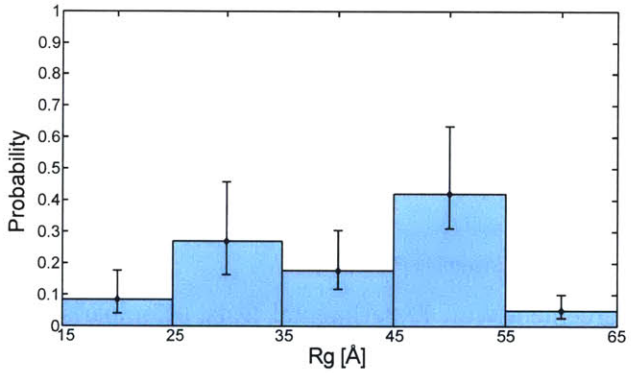

Figure III.3: Distribution of radii of gyration in the calculated ensemble. 95% confidence intervals show that the bar heights are significantly different from zero.

III.C.2 Residual Secondary Structure in α-Synuclein

To assess the secondary structure content in the Bayes ensemble, we used the STRIDE secondary structure assignment algorithm (Heinig and Frishman 2004), to calculate the propensity of each residue, in every structure in the ensemble, to adopt one of three mutually exclusive classes of secondary structure: helix, strand (also referred to as extended) and Other (see Methods). Analysis of individual structures within the ensemble reveals that the highest helical content is 20% while the highest strand content is approximately 28%. Of note, the highest weighted structures in the Bayes ensemble have helical content less than 15% and strand content less than 25% (Figure III.4). Nevertheless, the ensemble average secondary structure content is considerably less; i.e., the overall strand content is less than 11% and the helical content less than 2% (Table III.1). However, the ensemble average, which corresponds to the

experimentally observed value, is in excellent agreement with estimated secondary structure content obtained from CD spectroscopy (Rekas, Knott et al. 2010) (Table III.1). Although the experimental error bounds and the confidence intervals from the BW algorithm are relatively large for the helical and strand content, both BW and CD spectroscopy agree that the protein has minimal helical and strand content.

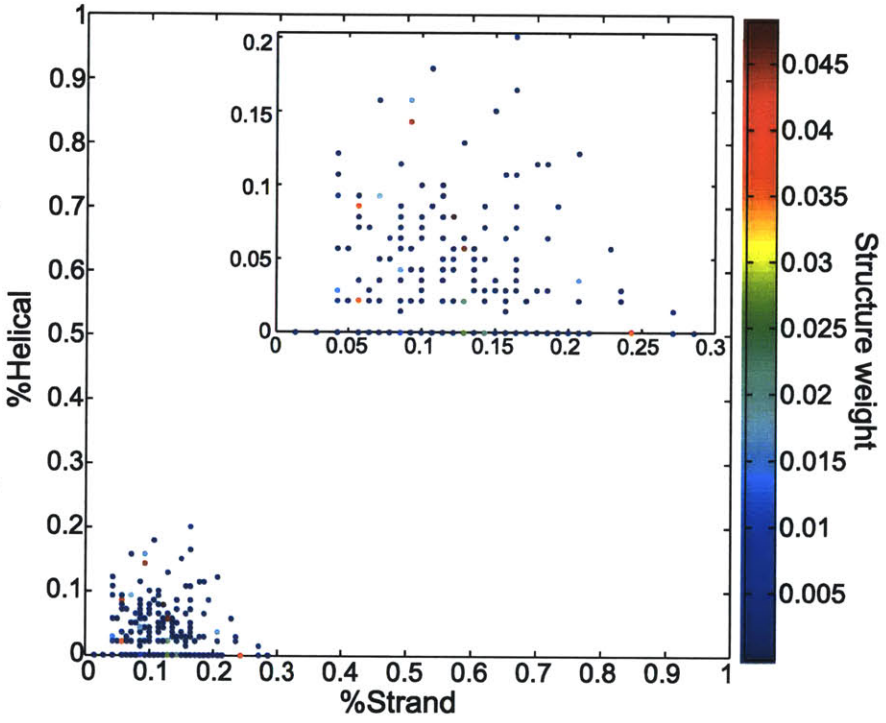

Figure III.4: Helical and Strand (or extended) content for each structure in the ensemble. The inset is an expanded view of the data.

	Ensemble Average	CD
Helix	0.03 (0.02-0.04)	0.02±0.03
Strand	0.11 (0.10-0.13)	0.11±0.07
Other	0.85 (0.84-0.87)	0.86±0.22

Table III.1: Ensemble Average Secondary Structure Content (with 95% confidence intervals) and Experimental values obtained from CD spectroscopy (with experimental error bounds).

In addition to the overall secondary content of the protein, we also computed the expected (or ensemble average) relative helix and strand propensities for each residue in the protein with their corresponding 95% confidence intervals (Figure III.5). On average, most of the helical propensity resides in residues 52-64. This region contains a highly conserved

hexamer motif within the 5th 11-mer imperfect repeat, which is proposed to form amphipathic α-helices (George, Jin et al. 1995, Davidson, Jonas et al. 1998). Additionally, within the NAC segment the strand propensity is peaked in the immediate vicinity of residue 78. Interestingly, this region, NAC(8-18), has been experimentally determined to be the minimal toxic aggregate forming segment in α-synuclein *in vitro* (El-Agnaf and Irvine 2002).

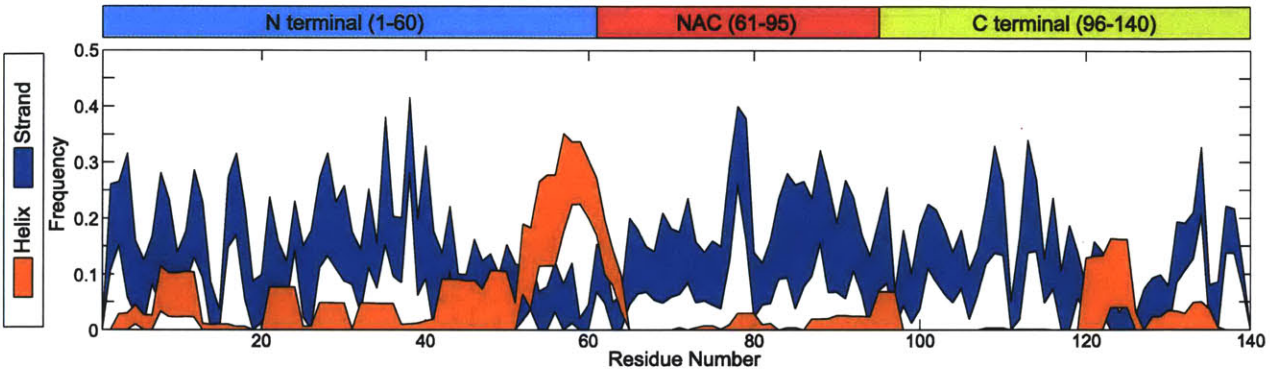

Figure III.5: Ensemble average secondary structure propensity. For each residue we present the probability to adopt a helical structure (orange area) vs. a strand structure (blue area). The thickness of the lines corresponds to the 95% confidence interval.

To further demonstrate that α-synculein samples structures with varying amounts of secondary structure, we explicitly show four conformations in Figure III.6. The N-terminal region is marked in blue, the NAC region in red and the C-terminal region in yellow. In Figure III.6A-C three structures are shown that contain varying degrees of helical content in the N-terminal and NAC region. Figure III.6A shows a structure containing a helix between residues 42-64; this helical conformation is in agreement with the contiguous helix model. Figure III.6B shows a structure that contains a helix in residues 74-82 in the middle of the NAC region and Figure III.6C presents two helices, one in the range 52-62 and the other in the range 15-24. Figure III.6D shows a structure with significant strand content in the NAC region – in particular residues 68-94.

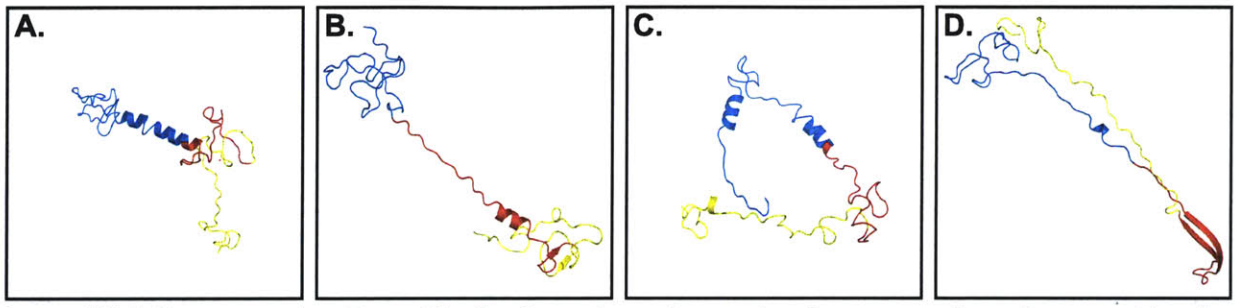

Figure III.6: Representative Ensemble Conformations. A sample of structures from the ensemble of α-synuclein. In all structures blue denotes the N terminal region (residues 1-60); red denotes the NAC region (residues 61-95) and yellow denotes C terminal region (residues 96-140).

III.C.3 Long Range Contacts in α-Synuclein

A number of studies have used Paramagnetic Relaxation Enhancement (PRE) experiments to detect long range contacts in α-synuclein (Bernado, Bertoncini et al. 2005, Dedmon, Lindorff-Larsen et al. 2005, Sung and Eliezer 2007, Koo, Choi et al. 2009, Rospigliosi, McClendon et al. 2009, Wu, Weinstock et al. 2009). These experiments allow for the detection of interactions between a paramagnetic group and nuclear spins of residues at a distance up to 25Å away (Gillespie and Shortle 1997). Some of these studies argue that long range contacts, especially involving the N-terminal (residues 1-60) and C-termini (residues 96-140), can be found in the unfolded ensemble of α-synuclein. To determine whether our data are consistent with these observations, we computed the distribution of such long range contacts the Bayes ensemble. For these calculations we define a long range contact between the N- and C-terminal regions to occur when the center of mass of the N-terminal region (residues 1-60) and the center of mass of the C-terminal region (residues 96-140) are within 25Å. We computed these center of mass distances for each structure and used these data to compute the distribution of such distances along with the associated confidence intervals (Figure III.7). Figure III.7 demonstrates that structures in the Bayes ensemble span a wide range of N- to C- terminal distances, ranging from less than 25Å to more than 125Å. In addition, a significant fraction (0.14 with a 95% confidence interval of 0.04-0.23) of the ensemble has structures that place the center of masses of the N- and C-terminal regions within 25Å of one another.

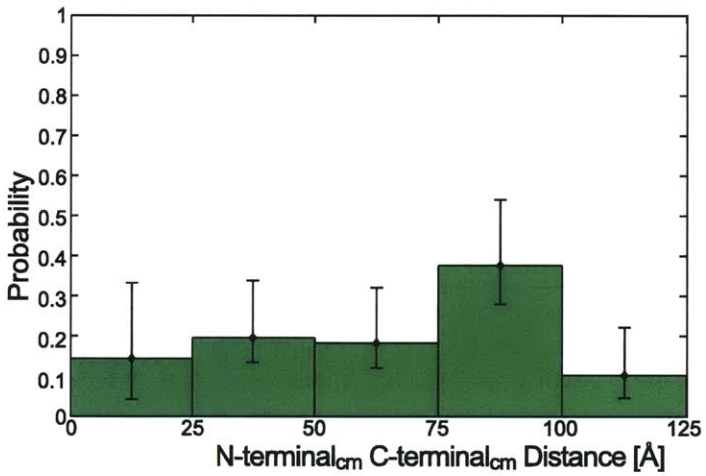

Figure III.7: Distribution of N- to C-terminal distances in the calculated ensemble. The x-axis represents the distance between the center of mass of the N-terminal region and the center of mass of the C-terminal region. 95% confidence intervals show that the bar heights are significantly different from zero.. Two low probability structures had N- to C-terminal distances higher than 125Å we therefore excluded this information from the distribution.

Since results from PRE experiments correspond to ensemble averages, we also computed a Residual Contact Map (which is a function of the ensemble average number of long range contacts per residue) to better compare our results to the previous PRE data (Figure III.8A). This pseudo-energy difference map represents the stability of a long range contact between two residues in the Bayes ensemble compared to what one would expect from a random coil ensemble (see Methods) (Dedmon, Lindorff-Larsen et al. 2005, Allison, Varnai et al. 2009). These data suggest that there is a distinct preference for forming long range contacts between the C terminal (residues ~120-140) and the N terminal region (residues 1-61) and also between the C terminal and the NAC region (residues ~61-70). In essence, residues 120-140 in the C-terminal region make contact with residues 1-70, which encompass both the N-terminal region and the beginning of the NAC segment. A less favorable contact forms between the NAC region (residues ~80-95) and the N terminal region (residues ~1-30); i.e., data are in qualitative agreement with prior experimental observations made from PRE data (Dedmon, Lindorff-Larsen et al. 2005, Allison, Varnai et al. 2009). In this regard, it is important to note again that our α-synuclein ensemble was generated without incorporating any data from prior PRE experiments and, therefore, no explicit distance constraints were used to construct the model.

In Figure III.8B, we show the most stable contact for each residue, along with the corresponding 95% confidence interval. The relatively small error bars for residues 20-70 and

residues 120-135 suggests that the model is relatively certain about these particular inter-residue contacts. However, the large error bars between residues 70-90 argues that the model is unsure about the inter-residue contacts in this region. These data complement the residual contact map in Figure III.8A; e.g., the residual contact map suggests that there is a relatively small preference for forming long range contacts between the NAC (residues ~80-95) and the N terminal region (residues ~1-30), however, the uncertainty analysis (Figure III.8B) suggests that the model is very uncertain about this particular observation.

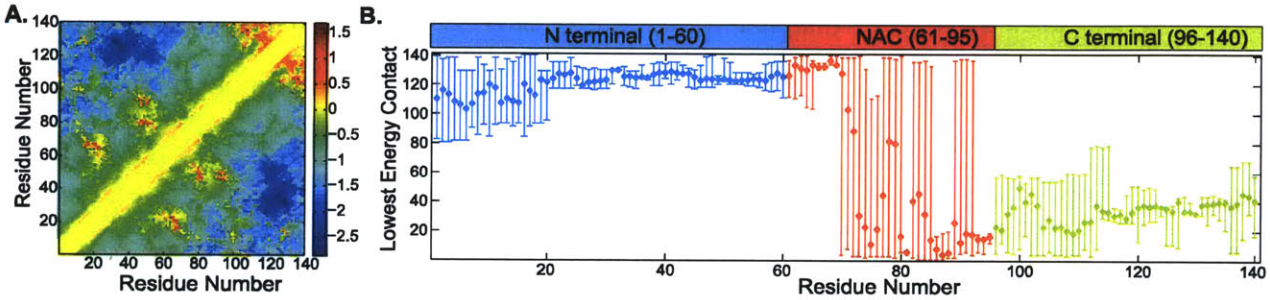

Figure III.8: Stabilized Long-Range Contacts. Pseudo-energy values are calculated for each contact as $-\ln\left(p_{ij}^{Ensemble} / p_{ij}^{RC}\right)$. A contact is defined between residues where the Cα atoms are less than 25Å apart. A large negative pseudo-energy (in blue) represents contacts that are energetically favorable in the ensemble compared to the random coil ensemble (in units of kT). Positive values (color range yellow to red) represent relatively unfavorable contacts. B. For each residue we calculated the contact that is associated with the lowest pseudo-energy along with the 95% confidence interval. The x-axis is the residue number and the y-axis is the position of the residue that forms the lowest energy contact. We used the following color code to depict the different regions of the protein, blue-N terminal region (residues 1-60); red-NAC region (residues 61-95) and yellow-C terminal region (residues 96-140).

III.C.4 Potential Aggregation Prone Conformers in α-Synuclein

Given that a relatively small segment of α-synuclein, consisting of residues 68-78, which is found in the NAC region (i.e., NAC(8-18)), was experimentally determined to be the minimal toxic aggregate-prone segment in α-synuclein *in vitro* (El-Agnaf and Irvine 2002), we explored the conformational preferences of this segment. Structures that place this segment in a solvent exposed and extended orientation may be more likely to form toxic aggregates containing cross-β structure.

Figure III.9 shows the normalized solvent accessible surface area (SASA) of the NAC(8-18) region versus the number of residues in that segment that are in an extended conformation, as identified by STRIDE (Heinig and Frishman 2004). Calculations of the SASA only included the

atoms H-N-Cα-C-O as this ensures that large SASA values identify structures that can form the intermolecular hydrogen bonds that are needed for cross-β structure formation. The Bayes ensemble contains several structures that place the aggregation prone segment, NAC(8-18), in a relatively extended and solvent exposed orientation. We define a residue as solvent exposed when it has a normalized SASA > 40%, as this cutoff has been used in previous studies and useful results were obtained (Stultz, White et al. 1993). In total, the fraction of structures that have the NAC(8-18) segment in a relatively extended and solvent exposed orientation is 0.08 with a 95% confidence interval 0.03-.12.

It has been postulated that the formation of long range contacts in α-synuclein may provide a mechanism to shield regions of the NAC segment (Dedmon, Lindorff-Larsen et al. 2005). Burying regions of the NAC segment could potentially hinder the formation of cross-β structure and the formation of toxic aggregates. To investigate the relationship between solvent exposure of the NAC(8-18) segment and long range contacts, we computed the SASA of the structures that have the center of mass of the N-terminal and C-terminal regions within 25Å. We find that the majority of structures that have the afore-mentioned long range contacts also place the NAC(8-18) segment in a solvent exposed orientation; i.e., 65% of structures (28%-100%, confidence interval) of structures with long range contacts have the NAC(8-18) segment with a SASA> 40%.

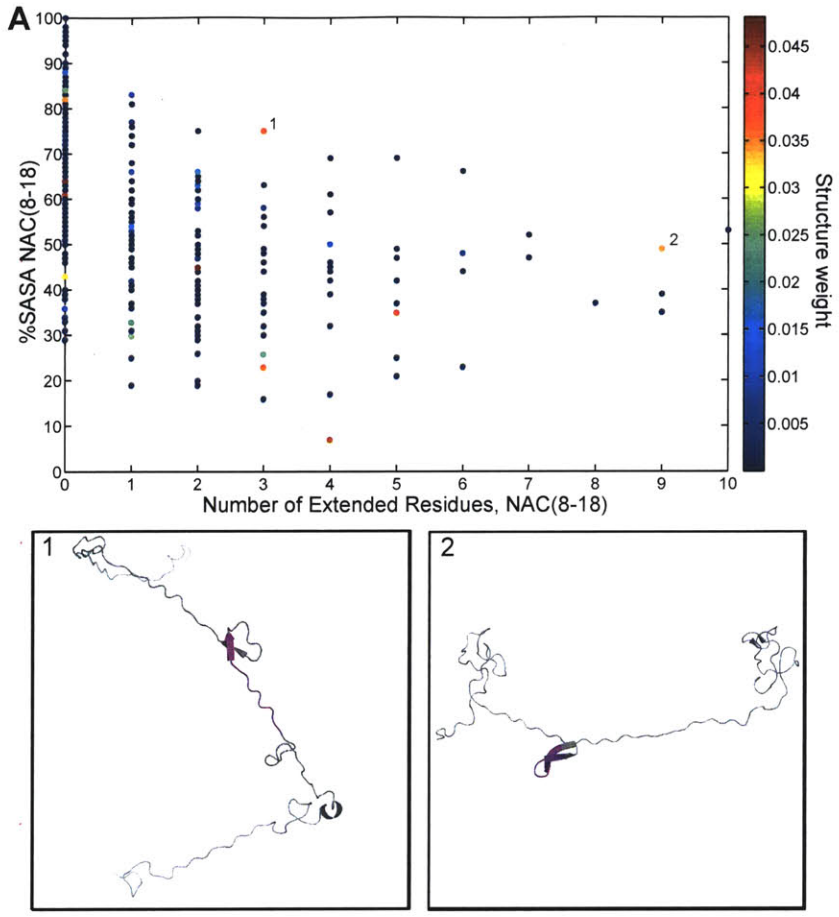

Figure III.9: The SASA vs. number of residues in an Extended orientation for the NAC(8-18) region. Two relatively high weighted structures with an exposed and a significant Extended content (more than 3 residues) for the NAC(8-18) segments are explicitly shown.

III.C.5 A Potential Mechanism for Helical Self-Association

In a recent study α-synuclein was isolated from human RBCs in a tetrameric form and the CD spectrum of this tetramer was quite distinct from that of recombinant α-synuclein obtained from *E. coli* (Bartels, Choi et al. 2011). Indeed, the spectrum of the tetramer had minima at 208 and 222 suggesting that, on average, the tetrameric structure had considerable helical structure.

Our data suggest that monomeric α-synuclein samples structures that have at most 20% helical content (Figure III.4). The structure with the highest helical content is shown in Figure III.10A. In general, the associated helix has a hydrophobic patch on one side (Figure III.10B), that is akin to hydrophobic faces that have been observed in other proteins that form helical bundles (Mathews, Bethge et al. 1979, Banner, Kokkinidis et al. 1987, Kamtekar and Hecht 1995). These data are consistent with a model where helical segments within structures in the

unfolded ensemble associate via hydrophobic interactions to form a tetrameric structure. If self-association of preformed helical structures was the dominant mechanism underlying the formation of tetrameric structures, then the expected helical content of the α-synuclein tetramers would be at most 20%. Interestingly, using the CD spectrum of the tetrameric structure (kindly provided by Tim Bartels and Dennis Selkoe), we obtain a predicted helical content of 29%, with an error of approximately 10%, using the program K2d (Andrade, Chacon et al. 1993, Merelo, Andrade et al. 1994).

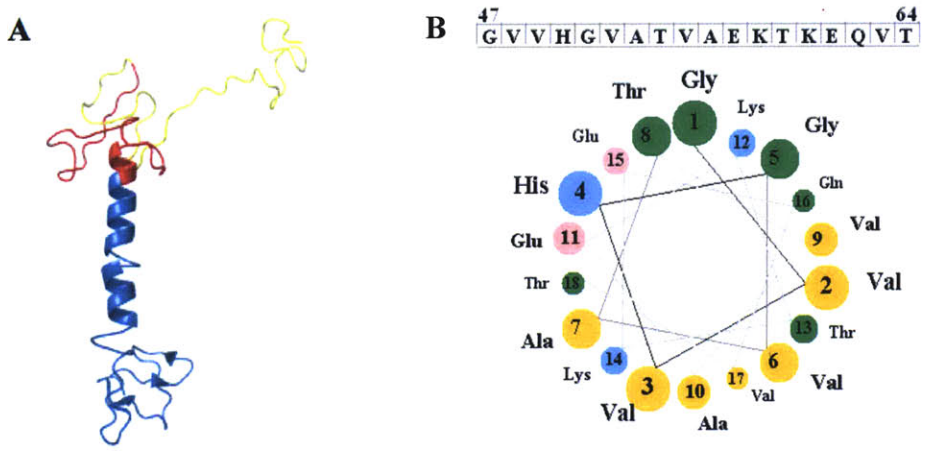

Figure III.10: A. Structure of the conformation within the ensemble that has the longest helical segment (expanded view of structure shown in Figure III.6A). As before, blue denotes the N terminal region (residues 1-60); red denotes the NAC region (residues 61-95) and yellow denotes C terminal region (residues 96-140). B. Associated helical wheel: orange - non polar residues orange, green - polar uncharged residues, blue - basic, pink - acidic.

More recently, Wang et al (Wang, Perovic et al. 2011) were able to obtain NMR data on a tetrameric form of α-synuclein that was purified from *E. coli*. Weak (i,i+3) Nuclear Overhauser Enhancements (NOEs) and secondary chemical shifts indicated helical propensity in residues 4-103. Intermolecular PREs, obtained using mixtures of α-synculein with and without the spin label, suggested that the tetramer forms by the association of amphipathic helices formed within the region consisting of residues 50-103. It is important to note, however, that the NMR data are not consistent with a fully folded helix. Instead, they suggest transient helical formation with an overall helical content of approximately 20% (Tom Pochapsky, personal communication) (Wang, Perovic et al. 2011). These data are qualitatively consistent with our model presented in

Fig. 10, which includes an amphipathic helix consisting of residues 47-64, and with our finding of α-helical propensity throughout the N-terminal region.

III.D Conclusion

Although monomeric α-synuclein is intrinsically disordered in solution it can adopt conformations that have varying secondary structure content depending on its environment (Uversky 2003). A number of spectroscopic studies suggested that WT-α-synuclein in the presence of membranes can form helical structures, and two types of helical configurations have been observed: a continuous extended helix and two antiparallel helices separated by a short linker (Bussell and Eliezer 2003, Chandra, Chen et al. 2003, Ulmer, Bax et al. 2005, Borbat, Ramlall et al. 2006, Georgieva, Ramlall et al. 2008, Jao, Hegde et al. 2008, Ferreon, Gambin et al. 2009, Trexler and Rhoades 2009, Georgieva, Ramlall et al. 2010). By contrast, α-synuclein was shown to acquire significant β sheet content when it self-associates *in vitro* (Apetri, Maiti et al. 2006). The most ordered form of these aggregates are fibrils which were found to contain cross-β-sheet structures (Serpell, Berriman et al. 2000). Most recently, an α-synuclein tetramer has been isolated from both human red blood cells and *E. coli* and it has been argued that this structure is the dominant form under physiological conditions (Bartels, Choi et al. 2011, Wang, Perovic et al. 2011). While initial reports suggested this structure was a 'folded tetramer' rather than an 'unfolded monomer', it is important to note that the data from recent NMR studies suggest that the actual amount of structural order is fairly low in the tetramer and that there is only fractional helix formation (Wang, Perovic et al. 2011). These observations highlight the need to obtain an accurate structural ensemble that describes the accessible states of the protein.

While a number of studies have generated ensembles for the unfolded state of α-synuclein (Bernado, Bertoncini et al. 2005, Dedmon, Lindorff-Larsen et al. 2005, Allison, Varnai et al. 2009, Koo, Choi et al. 2009, Wu, Weinstock et al. 2009), the majority of these ensembles has not provided a detailed analysis of residual secondary content within the ensemble and have not addressed recent data on what has been described as a physiologically dominant helical tetramer (Bartels, Choi et al. 2011). More importantly, existing studies have led to contradictory observations. Although one prior study suggested a small preference for residues 6-37 (in the N

terminal) to form helices compared to residues 103-140 (in the C terminal) (Allison, Varnai et al. 2009), another study did not find significant helical structure within the ensemble (Wu, Weinstock et al. 2009). Therefore, to further explore the nature of the unfolded state of α-synuclein and to understand the differential binding characteristics of the protein, we constructed and analyzed an ensemble that represents the unfolded state of monomeric α-synuclein in solution.

As we have previously noted, the construction of models that adequately represent the unfolded state of a protein is inherently difficult for a number of reasons. First, there is the conformational sampling problem; i.e., sampling all possible conformations of even a modestly sized protein is intractable. Nevertheless, the form of the underlying energy surfaces helps because the space of energetically favorable conformations is likely far less than the space of all possible conformations. Some studies have shown that straightforward Boltzmann sampling for some IDPs yield calculated observables that are in reasonable agreement with experiment, thereby suggesting that extensive sampling, by itself, is a plausible approach for generating representative ensembles for IDPs (Fawzi, Phillips et al. 2008, Sgourakis, Merced-Serrano et al. 2011). However, while it is reasonable to apply such a direct sampling approach to relatively small proteins, the prospect of extensively sampling the relevant conformational space of a protein that is 140 residues in length (at 300K) is daunting. In this regard, a number of approaches that do not rely on direct Boltzmann sampling of large proteins have been developed and useful insights have been obtained using these methods (Feldman and Hogue 2000, Bernadó, Blanchard et al. 2005, Jha, Colubri et al. 2005, Marsh and Forman-Kay 2009).

In the present study we use a fragment based approach to sample energetically favorable conformations of the entire 140 residue protein. The motivation for this approach arose from a prior study that demonstrated that sampling the conformational states of fragments from folded proteins may reproduce the backbone structure of that peptide's structure in the context of the entire protein (Ho and Dill 2006). In our method, the protein was divided into eight residue long overlapping segments and REMD was used to sample the conformational space of each peptide. Eight residue segments were chosen because this length corresponds, roughly, to the average persistence length of a polypeptide (Jha, Colubri et al. 2005). Structures for α-synuclein were generated by combining these overlapping segments, and subsequent energy minimization of each reconstructed conformer ensures that each structure corresponds to local energy minima on

the potential energy surface of the protein. While the approach is computationally efficient, we recognize that focusing on sampling small peptides may limit the formation of long range interactions within the final ensemble. To mitigate this, once the peptide fragments have been combined to generate the α-synuclein sequence, the entire protein is energy minimized thereby allowing the different peptides to "see" with one another. Using the resulting structural library with the BW algorithm we arrive at a Bayes ensemble that: 1) contains conformations that correspond to local energy minima on the potential energy surface and 2) agrees with the available experimental data. Nevertheless, while our approach is computationally very efficient, we recognize that other sampling approaches for the initial structural library (e.g. using different lengths for the peptide segments) may lead to different structural libraries and this fact introduces some uncertainty in our analysis.

One additional source of uncertainty is the inherent degeneracy of the problem of constructing a good ensemble, even after the precise structural library has been specified. Given that the number of degrees of freedom (i.e., the number of energetically favorable conformations) is typically much greater than the number of independent experimental observables, the problem of choosing, or weighting, a set of structures is inherently degenerate; i.e., there are many possible ways of weighting the structures that will agree with any given set of experimental observations (Fisher, Huang et al. 2010).

To deal with these sources of uncertainty, the BW method calculates a probability distribution over the space of ensembles. This posterior density function naturally leads to a new metric, the uncertainty parameter, which quantifies our uncertainty in the Bayes ensemble. This uncertainty parameter is akin to the standard deviation in a Gaussian distribution and reflects the overall spread of the calculated this probability distribution. The uncertainty parameter varies between 0 and 1 where a value of 0 suggests that the Bayes ensemble is correct. By contrast when the uncertainty parameter is non-zero one cannot be certain that the Bayes ensemble is correct. However, in this latter instance one can express values with the appropriate confidence intervals (Fisher, Huang et al. 2010). Therefore, the method provides a rigorous means to quantify the overall uncertainty in the final results.

The BW algorithm yields a Bayes ensemble that is in agreement with data obtained using NMR chemical shifts (Rao, Kim et al. 2009), RDCs (Bertoncini, Fernandez et al. 2005) and the radius of gyration as determined by SAXS experiments (Binolfi, Rasia et al. 2006). Surprisingly

we find that some conformers in the ensemble are nearly as compact as a folded globular protein with the same amino-acid length. In addition, the Bayes ensemble contains structures that have a radius of gyration that is larger than the average radius of gyration that would be expected from a 140 residue random coil. This highlights the fact that a single experimental value for the radius of gyration provides little insight into the full range of conformations that the protein can adopt in solution. Two recent experimental studies suggested the existence of distinct classes of conformers describing α-synuclein equilibrium (Sandal, Valle et al. 2008, Frimpong, Abzalimov et al. 2010). Both studies argue that α-synuclein contains a range of conformations, where some are quite compact and others are quite extended and random-coil like. Our data are in agreement with these observations and quantify the extent range of radii of gyration within the ensemble. Moreover, these studies highlight the fact that while the overall secondary structure content of the ensemble is negligible (about 7% of the population was suggested to contain "β-like" conformation), there are subpopulations of structures that have more significant helical and strand content – a finding in agreement with our observations.

The resulting structural ensemble provides additional insight into the secondary structure propensities within different regions of the protein. We find that the Bayes ensemble contains several structures that have helical segments of varying length in the N-terminal region, extending into the NAC segment. In total, the helical regions span residues 1-92; i.e., the segment that has been shown to adopt either a continuous helix or a broken helix in the presence of lipid membranes (Ferreon, Gambin et al. 2009, Trexler and Rhoades 2009, Georgieva, Ramlall et al. 2010). These data are consistent with a model of lipid binding where interaction with the membrane stabilizes these helical segments leading to the formation of either a continuous or a broken helix, depending on the precise experimental conditions. In this sense, the presence of relatively short helical segments may serve as intermediates that enable fast and efficient binding to lipid membranes. These data are of particular importance because interactions between α-synuclein, in its helical form, and membranes may play a role in cellular dysfunction in patients with Parkinson's disease (Auluck, Caraveo et al. 2010).

By contrast, on average, significant probability for extended structure is found throughout the α-synuclein sequence. These data are in qualitative agreement with previous Raman spectroscopic studies that suggest that the protein adopts an ensemble of rapidly interconverting secondary structural elements (Frishman and Argos 1995). Of particular interest is the region

spanning residues 68-78 in the NAC segment because this has been shown to be the minimal toxic peptide that is also can initiate α-synuclein aggregation *in vitro* (El-Agnaf and Irvine 2002). The probability of extended structure is relatively peaked around this region, thereby suggesting that this segment has an intrinsic predisposition to form extended structure that may initiate the formation of β-sheet rich aggregates. However, in order to form intermolecular hydrogen bonds with other α-synuclein molecules, this segment must be exposed to solvent. Therefore, structures that place this segment in a solvent exposed and extended conformation may be more prone to form toxic aggregates. Our analysis suggests that approximately 8%, with a 95% confidence interval of 3-12%, of the structures in the ensemble have the NAC(8-18) segment in an extended and solvent exposed orientation. This suggests that the unfolded ensemble of α-synuclein contains preformed conformations that can readily form β-sheet rich toxic aggregates.

In addition to these insights, our data further clarify the role of long range contacts in the protein. Previous studies that have constructed ensembles based on the results of PRE experiments have found conflicting findings, even though many of these experiments were performed under similar experimental conditions. One study suggested that long-range interactions occur between residues 85-95 of the NAC and the C terminal region (specifically residues 110-130) (Bernado, Bertoncini et al. 2005). Other studies suggested the formation of long range contacts between the highly charged C-terminus (residues 120-140) and the large hydrophobic center (residues 30-100) resulting in a hydrodynamic radius significantly smaller than that expected for a random coil structure (Dedmon, Lindorff-Larsen et al. 2005, Allison, Varnai et al. 2009). In another study, done under similar conditions, it was suggested that long range contacts form between the N-terminus and the NAC region in contrast to the previously mentioned studies (Wu, Weinstock et al. 2009). Therefore, while these data have provided new insights into the nature of the unfolded state of α-synuclein in solution, they leave the precise role of any long range interactions in the protein unclear.

Our data suggest that, on average, there are long range contacts between the N and C termini of the molecule. Interestingly, the Bayesian estimates allow us to say with confidence that the N-terminal region and the first nine residues from the N terminal portion of the NAC make, on average, contacts with the C-terminal region of the protein – a result that is in qualitative agreement with prior studies (Dedmon, Lindorff-Larsen et al. 2005, Allison, Varnai et al. 2009). It has been suggested that these long range interactions provide a mechanism that

effectively shields the aggregation prone region, and thereby minimizes the extent of aggregation (Bertoncini, Jung et al. 2005, Dedmon, Lindorff-Larsen et al. 2005, Allison, Varnai et al. 2009). However, a more detailed look at the actual distribution of structures within the ensemble (as opposed to an analysis of the ensemble average data) finds that most of the structures in our ensemble that contain long range contacts between the N- and C-termini also place the NAC(8-18) segment in a solvent exposed conformation. An example of one such structure is shown in Figure III.6D. Consequently, it is not clear that separation of the N- and C- termini is required to expose the most aggregation prone regions of the sequence. This claim is supported by recent PRE experiments comparing WT α-synuclein and A30P, E46K and A53T naturally occurring mutants, which were all shown to have a higher aggregation rate *in vitro*. Results of this study suggest that A30P and A53T mutants did not have a significant decrease in the N– and C-termini contacts. Moreover, E46K presented an increase in these long range contacts (Rospigliosi, McClendon et al. 2009). These data bring into question whether long range contacts play a key role in regulating aggregation of α-synuclein.

We note that our model did not use any PRE derived distance restraints. While PRE-derived data have provided valuable information into the presence or absence of long range contacts in several IDPs, it requires introducing a paramagnetic probe into the protein (Mittag and Forman-Kay 2007). However, it may be that such probes alter the accessible states of the unmodified protein. In light of these observations, and the fact that some of the PRE-derived results are contradictory, we did not explicitly use PRE-derived data when building our ensemble. Nevertheless, we obtain results that corroborate and clarify many aspects of the prior PRE studies.

In addition, we recognize that there may be additional contacts between the N-terminal, NAC, and C-terminal regions, but we cannot make statements about these interactions with confidence given the very wide error bars associated with residues in the more central region of the NAC segment (Figure III.8B). Interestingly, our uncertainty in the precise contacts that involve the entire NAC region is also reflected in the literature as the NAC region is suggested to interact with the C terminal in some studies while other studies suggest that it interacts with the N terminal instead (Bertoncini, Jung et al. 2005, Dedmon, Lindorff-Larsen et al. 2005, Sung and Eliezer 2007, Cho, Nodet et al. 2009, Rospigliosi, McClendon et al. 2009, Wu, Weinstock et al. 2009). In short, our model is unable to distinguish between these two possibilities with certainty.

Recent data suggest that α-synuclein forms helical tetramers under physiological conditions (Bartels, Choi et al. 2011, Wang, Perovic et al. 2011). Our data suggest that monomeric α-synuclein samples structures that have at most 20% helical content and that some of these helices are amphipathic in character. These data are consistent with a model where tetrameric structures are formed via the interaction of hydrophobic patches on these amphipathic helices. Indeed there are many examples in the literature of such four-helical bundle structures composed of amphipathic helices (Mathews, Bethge et al. 1979, Banner, Kokkinidis et al. 1987, Kamtekar and Hecht 1995). Wang et al. independently proposed a model for the tetrameric state of α-synuclein on the basis of NMR data in which transiently formed amphipathic helices interact in just such a manner (Wang, Perovic et al. 2011).

Our results argue that the unfolded state of α-synuclein contains a heterogeneous set of conformations of both highly compact and extended structures, and that while the overall secondary structure content of these structures are low, there are regions that have a relatively high propensity for helical and extended structure. Regions with a significant propensity for either helical or strand content may facilitate the formation of lipid-associated helical structures, helical tetrameric structures, and aggregates that are rich in β-sheets. Our results also provide quantitative estimates for the percentage of structures that are compact, have long-range contacts between the N- and C-termini, and that have the minimal toxic aggregation fragment of α-synuclein that is in a position that is poised to make intermolecular β strands. In sum, these data provide a comprehensive view of the unfolded ensemble of monomeric α-synuclein in solution and explains how different ordered structures conformers can arise from this disordered protein.

III.E Methods

III.E.1 Generation of an α–Synuclein Structural Library

The sequence of α–synuclein was divided into eight residue long segments resulting in 28 segments in total (the C-terminal segment was five residues long). Each segment had three residues overlap with the adjacent segments. A similar protocol was used to describe K18, an intrinsically disordered protein of comparable size, 130 amino acids long (Fisher, Huang et al. 2010). The size of the segments was chosen based on the average persistence length of a

polypeptide (Jha, Colubri et al. 2005). Conformations for segments of this length were shown to be successfully sampled using a replica exchange molecular dynamics (REMD) (Sugita and Okamoto 1999) procedure (Fink 2006).

Each segment underwent REMD with the EEF1 (Lazaridis and Karplus 1999) implicit solvent model using CHARMM (Brooks, Bruccoleri et al. 1983). A total of 16 replicas, each at a different temperature were used. Temperatures were spaced exponentially in the range 280 to 700K. Segments were run for 10ns, and structures were collected from the last 5ns of the 298K heat bath, allowing 5ns of equilibration period. A total of 5000 conformations per segment were collected.

Full length α–Synuclein conformations were generated by piecing together the segments one at a time, starting with the N-terminal segment. Each segment was clustered according to the three overlapping residues at its ends. The segment to be added to the growing polypeptide chains was chosen from the cluster that had the most structural similarity in the overplaying region. The first residue coordinates of the overlapping segments were taken from the C terminal of the one segment and the two others from the N terminal of the adjoining segment. At the end of the procedure the full length structure was subjected to 1,000 steps of steepest descent minimization followed by 10,000 steps of adopted basis Newton-Raphson minimization to relieve any bad contacts in the molecule. Only structures with a negative energy were chosen for the structural library. Following the process we found the structural library generated was composed of structures that were mainly compact when comparing their radius of gyration (Rg) to the one obtained by SAXS experiments. Therefore the combined pre-energy minimization structures were used in additional energy minimization using an Rg restraint. Rg restraints varied from 27Å to 75Å. This process ensures that a wide range of conformations were generated. At the end of the process ~100,000 structures were generated.

The structural library was reduced in size to 299 structures using our previously described pruning algorithm (Fisher, Huang et al. 2010). This number of structures was shown to be able to provide a good model for the K18 tau segment of comparable size (130 residues) (Fisher, Huang et al. 2010).

III.E.2 Generation of an α–Synuclein Ensemble from the Pruned Structural Library and the Calculation of Confidence Intervals

In order to obtain the sets of weights for the pruned structural library, we employed the BW algorithm as previously described (Fisher, Huang et al. 2010). In this method, one generates a posterior distribution which represents the probability of all possible weighting schemes over the 299 structures, given the available experimental data. Experimental measurements used were C, Cα, Cβ and N chemical shifts (Rao, Kim et al. 2009), N-H RDCs (Bertoncini, Fernandez et al. 2005) and radius of gyration (Binolfi, Rasia et al. 2006). The carbonyl chemical shift value for residue 140 from the set of experimental data points was not used, as it was an extreme outlier from the other data (Rao, Kim et al. 2009). To implement the BW method we first need to calculate the corresponding chemical shifts for each atom, along with RDCs and the radius of gyration, in each structure. Chemical shifts were calculated with SHIFTX (Neal, Nip et al. 2003), and the radii of gyration were calculated with CHARMM (Brooks, Bruccoleri et al. 1983). The RDCs of each individual conformer in the ensemble was calculated with PALES (Zweckstetter and Bax 2000) based on a 'global alignment' model, i.e. using the entire protein structure. This is in contrast to a 'local alignment' model, in which the RDCs are calculated from short segments of the protein. It has been suggested that one can reproduce experimental RDCs with a smaller number of conformers when using a local alignment model as compared to a global alignment model (Marsh, Baker et al. 2008, Nodet, Salmon et al. 2009); nevertheless, we were able to obtain good agreement with the experimental RDCs using the global alignment method with a relatively small number of highly populated conformers.

The BW algorithm incorporates information from both the experimental errors and the errors associated with predictions for the experimental values of interest (Fisher, Huang et al. 2010). Experimental errors were taken to be 0.3ppm (chemical shifts), 1Hz (RDCs) and 2Å (Radius of Gyration), respectively. As prediction errors for chemical shift values have been rather extensively studied, they were also included in the expression for the posterior distribution (Neal, Nip et al. 2003, Fisher, Huang et al. 2010).

Here, we provide a very brief review of the theoretical aspects of the BW framework; for a comprehensive description see Fisher et al (Fisher, Huang et al. 2010). Formally, the posterior probability distribution conditioned on the observed experimental data is obtained from Bayes' rule:

$$f_{\vec{W}|\vec{M}}(\vec{w}|\vec{m}) = \frac{f_{\vec{M}|\vec{W}}(\vec{m}|\vec{w}) f_{\vec{W}}(\vec{w})}{f_{\vec{M}}(\vec{m})} \tag{1}$$

where $f_{\vec{W}}(\vec{w})$ is a the prior probability distribution – for brevity, the specific form will not be reproduced here – and $f_{\vec{M}|\vec{W}}(\vec{m}|\vec{w})$ is the likelihood function for the vector of experimental observations, $\vec{m}$. We assume that the likelihood function can be decomposed as $f_{\vec{M}|\vec{W}}(\vec{m}|\vec{w}) = f^{R_g}_{\vec{M}|\vec{W}}(m^{R_g}|\vec{w}) f^{RDC}_{\vec{M}|\vec{W}}(\vec{m}^{RDC}|\vec{w}) f^{CS}_{\vec{M}|\vec{W}}(\vec{m}^{CS}|\vec{w})$ where each of the components is (multivariate)-Gaussian. Specifically, the likelihood functions are:

$$f^{R_g}_{\vec{M}|\vec{W}}(m|\vec{w}) = \left[2\pi\varepsilon^2_{R_g}\right]^{-1/2} exp\left[-\frac{\left(m - E_{R_g}[m|\vec{w}]\right)^2}{2\varepsilon^2_{R_g}}\right]$$

$$f^{RDC}_{\vec{M}|\vec{W}}(\vec{m}|\vec{w}) \propto \int_{-\infty}^{\infty} \prod_{i=1}^{N_{RDC}} \left(2\pi\varepsilon^2_{i,RDC}\right)^{-1/2} exp\left[-\frac{\left(m_i - \lambda E_{R_g}[m_i|\vec{w}]\right)^2}{2\varepsilon^2_{i,RDC}}\right] d\lambda \tag{2}$$

$$f^{CS}_{\vec{M}|\vec{W}}(\vec{m}|\vec{w}) = \prod_{i=1}^{N_{CS}} \left[2\pi\left(\varepsilon^2_{i,CS} + \alpha^2_{i,CS}\right)\right]^{-1/2} exp\left[-\frac{(m_i - E_{CS}[m_i|\vec{w}])^2}{2\left(\varepsilon^2_{i,CS} + \alpha^2_{i,CS}\right)}\right]$$

Here, the letter ε denotes an experimental error, α denotes a prediction error and λ is a factor for uniformly scaling the RDCs to account for uncertainty in the magnitude of alignment.

The Bayes estimate for the weight for each structure corresponds to the expected (or average) value of that structure's weight over the posterior distribution; i.e. $w^B_j \equiv \langle w_j \rangle_{BW} = \int d\vec{w} w_j f_{\vec{W}|\vec{M}}(\vec{w}|\vec{m})$. The uncertainty parameter is the average distance from the Bayes weights, or $\sigma_{\vec{w}^B} \equiv \left[\int d\vec{w} \Omega^2(\vec{w}^B, \vec{w}) f_{\vec{W}|\vec{M}}(\vec{w}|\vec{m})\right]^{1/2}$, where $\Omega^2(\vec{w}^B, \vec{w})$ is metric on the space of weight vectors called the Jensen-Shannon divergence (Fisher, Huang et al. 2010). To calculate these expected values, samples are taken from the posterior distribution using a Monte Carlo algorithm with Gibbs Sampling (Fisher, Huang et al. 2010). Each sample corresponds to a different weighting scheme over the 299 structures. 100 million samples were generated as an equilibration period for the Markov Chain generated from the Monte Carlo algorithm. This was followed by an additional 1 billion samples, which constitutes the "production run". We followed the running average of the posterior divergence to ensure that convergence was reached. To calculate the Bayesian averages and the associated confidence intervals, we used 50,000 equally spaced samples from the 1 billion samples; this reduces the overall computation time. For a given quantity (e.g., the expected solvent exposure of a given residue), we computed

this quantity using the chosen samples, yielding 50,000 estimates for the value of interest. The 95% confidence interval was obtained by finding the lower bound that excluded the bottom 2.5% of the estimates and the upper bound that excluded the top 2.5% of the estimates.

III.E.3 Random Coil Ensemble

The Residual Contact Map shown in Figure III.8A represents the stability of a long range contact between two residues in our ensemble compared to what one would expect from a random coil ensemble. Therefore to compute the contact map, we first need to generate a random coil ensemble for α-synuclein. We used the publicly available random coil ensemble posted in a web repository (Jha, Colubri et al. 2005). The ensemble contains 5000 structures; we therefore randomly selected 299 structures to insure the two ensembles are of the same size. The selection process was repeated 20 times in order to reflect the full ensemble and each measurement of interest was averaged over this collection. The random coil model used to form this ensemble uses statistical potential and excluded volume constraints (Jha, Colubri et al. 2005), no α-synuclein experimental data was included in generating the ensemble.

III.E.4 Secondary Structure Assignments

We clustered STRIDE results of α-helix, pi-helix and 3-10 helix into a super class that we refer to as Helix. In addition we cluster isolated bridge and extended results into a second super class we named Extended (or Strand). All other secondary structure assignments were combined into a single class denoted as Other. In order to have consistent definitions when comparing to the results obtained from CD spectroscopy (which assigned a helix, strand, turn and unstructured) (Rekas, Knott et al. 2010), we grouped the "turn" and the "unstructured" assignments into one category called "Other".

III.E.5 Solvent Accessible Surface Calculations

The solvent exposure surface area for each conformation was calculated using CHARMM (Brooks, Bruccoleri et al. 1983). The SASA of the entire protein was computed, but only data from the solvent exposure of the backbone atoms N-H-C-Cα-O were used, since these represent atoms that are essential for the formation of cross-β sheet interactions. The calculated SASA was normalized by dividing by solvent accessible surface of the backbone atoms when α-synuclein is in a fully extended conformation. A residue is said to be solvent exposed when its

normalized SASA > 40%, as this cutoff has been used in previous studies and useful results were obtained (Stultz, White et al. 1993).

III.E.6 Calculating Distributions of Ensemble Properties

Two plots were generated presenting probabilities calculated from the posterior distribution. The radius of gyration for each structure in the ensemble is calculated in CHARMM (Brooks, Bruccoleri et al. 1983) using the N-C-Cα atoms, structures are binned together in bins of 10Å. Summation of the structures probabilities (their weights) in each bin comprises the probability of that bin. 95% confidence intervals were then obtained using the 50,000 samples from the posterior distribution as outlined above. The histogram of N-terminal center-of-mass to C-terminal center-of-mass distances was generated in a similar fashion. Distances were calculated in CHARMM (Brooks, Bruccoleri et al. 1983) using the center of mass of N-C-Cα backbone atoms for residues 1-60, the N-terminal, and the center of mass of N-C-Cα backbone atoms for residues 96-140 – the C terminal. Structures we binned in bins of 25Å, corresponding to the maximal distance defined for formation of long range contacts and again the 95% confidence intervals were then obtained from the 50,000 samples from the posterior distribution.

III.E.7 Helical wheel diagram

To generate the helical wheel, we used the freely available Helical Wheel program (http://cti.itc.virginia.edu/~cmg/Demo/wheel/wheelApp.html). Amino-acid sequences taken from the conformation with the longest continuous helical structure were input to the Helical Wheel program to generate the associated diagram.

Chapter IV

The Dynamic Structure of α-Synuclein Multimers

IV.A Abstract

α-Synuclein, a protein that forms ordered aggregates in the brains of patients with Parkinson's disease, is intrinsically disordered in the monomeric state. Several studies, however, suggest that it can form soluble multimers *in vivo* that have significant secondary structure content. A number of studies demonstrate that α-synuclein can form β-strand rich oligomers that are neurotoxic, and recent observations argue for the existence of soluble helical tetrameric species *in cellulo* that do not form toxic aggregates. To gain further insight into the different types of multimeric states that this protein can adopt we generated an ensemble for an α-synuclein construct that contains a 10 residue N-terminal extension, which forms multimers when isolated from *E. coli*. Data from NMR chemical shifts and residual dipolar couplings were used to guide the construction of the ensemble. Our data suggest that the dominant state of this ensemble is a disordered monomer, complemented by a small fraction of helical trimers and tetramers. Interestingly, the ensemble also contains trimeric and tetrameric oligomers that are rich in β-strand content. These data help to reconcile seemingly contradictory observations that indicate the presence of a helical tetramer *in cellulo* on the one hand, and a disordered monomer on the other. Furthermore, our findings are consistent with the notion that the helical tetrameric state provides a mechanism for storing α-synuclein when the protein concentration is high; thereby preventing non-membrane bound monomers from aggregating.

This chapter was published in similar form in: Gurry, T.*, Ullman, O.*, Fisher, C.k., and Stultz, C.M. The Dynamic Structure of α-Synuclein Multimers. *J. Am. Chem. Soc.* **2013**, *135*, 3865-3872 (*Both authors contributed equally to this work)

IV.B Introduction

α-Synuclein is a 140-residue protein that has been implicated in the pathogenesis of a number of neurodegenerative diseases, collectively known as synucleinopathies, the most well-known of which is Parkinson's disease (Bellucci, Zaltieri et al. 2012). The most notable pathological characteristic of these diseases is the aggregation of α-synuclein into amyloid fibrils, which have significant β-sheet secondary structure (Spillantini, Schmidt et al. 1997, Uversky, Li et al. 2001). Although there is disagreement regarding whether the soluble oligomeric aggregates or insoluble aggregates are the most neurotoxic species, it is clear that α-synuclein self-association plays an integral role in neuronal dysfunction and death (Conway, Lee et al. 2000, Bucciantini, Giannoni et al. 2002, Kayed, Head et al. 2003, Danzer, Haasen et al. 2007, Winner, Jappelli et al. 2011). Given the importance of this protein in these neurodegenerative disorders, studies that help to elucidate its structure are of paramount importance.

However, the conformational landscape of α-synuclein is notoriously difficult to study, earning it the moniker of 'chameleon' due to its tendency to adopt different conformations under different experimental conditions (Uversky 2003, Drescher, Huber et al. 2012). This has led to seemingly contradictory data about the dominant putative states in solution versus those under physiologic conditions (Bartels, Choi et al. 2011, Wang, Perovic et al. 2011, Fauvet, Kamdem et al. 2012). While it is clear that monomeric α-synuclein is an intrinsically disordered protein (Weinreb, Zhen et al. 1996) in solution, recent data suggests that it can adopt a tetrameric state that has a relatively high helical content under physiologic conditions (Bartels, Choi et al. 2011, Wang, Perovic et al. 2011, Trexler and Rhoades 2012). By contrast, others have suggested that α-synuclein retains its monomeric disordered state *in cellulo* (Binolfi, Theillet et al. 2012, Fauvet, Kamdem et al. 2012).

Recently, NMR studies on an α-synuclein construct isolated from *E. coli*, which contains a 10 residue N-terminal extension, suggested that the protein can exist as a "dynamic tetramer" (Wang, Perovic et al. 2011). In short, these data are consistent with a model where the protein rapidly interconverts between different conformers, where some of these conformations are multimeric structures (trimers and tetramers) that contain significant helical content. To obtain a more comprehensive view of the types of structures that this particular α-synuclein construct can

adopt, we generated an atomistic model for α-synuclein in its multimeric form. While we recognize that it is not possible to capture all possible monomeric and multimeric conformations that this protein can adopt in solution, our hope was to build a low-resolution description of the dominant states of the protein. More precisely, we define a conformational ensemble to consist of a structural library $S = \{\vec{s}_i\}_{i=1}^{n}$, where $\vec{s}_i$ is the Cartesian coordinates of structure i, and a corresponding set of weights $\vec{w} = \{w_i\}_{i=1}^{n}$, where w_i is the population weight of structure i. In this sense, the number of structures in the ensemble, n, is a function of the resolution with which one wishes to view the conformational landscape of the system.

As prior studies on this construct suggest that the purified protein contains primarily monomers, trimers and tetramers, we focused on these specific forms for our ensemble (Wang, Perovic et al. 2011). Since we had previously constructed an ensemble for monomeric α-synuclein using NMR chemical shifts, RDCs and SAXS data (Ullman, Fisher et al. 2011), we used these structures to represent the disordered, monomeric fraction. Using NMR chemical shifts and NH RDCs obtained on an α-synuclein construct, which contains a 10 residue N-terminal extension, we determine the relative fractions of different multimeric forms within the ensemble.

IV.C Results and Discussion

To generate a set of energetically favorable multimers for the ensemble, we began with a set of "seed" structures that served as starting points from which a diverse library of multimeric structures could be built. Our previous study on α-synuclein suggested that the monomeric protein can sample amphipathic helices, which could in principle self-associate to form higher order structures (Ullman, Fisher et al. 2011). Hence, we constructed trimeric and tetrameric structures using amphipathic helices from the monomeric ensemble. Structures for both the trimeric and tetrameric species were obtained by threading these amphipathic helices onto three- and four-helix bundles, respectively, from the Protein Data Bank (PDB) such that the hydrophobic faces of these helices form the contact-interface (see Methods). A second helical tetrameric model was constructed using the available NMR data (Wang, Perovic et al. 2011). The model derived from the NMR data was obtained from a limited set of NOEs because a high

degree of spectral overlap is present even in three-dimensional data sets. Consequently, the resulting model is not intended to represent a "high-resolution" structure of the helical tetramer. Instead, it is a model, constructed from limited experimental data, which serves as a starting point for additional simulations. Indeed, all seed structures represent initial structures (derived from experimental data and from prior studies on the monomeric state) from which to begin sampling, rather than high-resolution structures for trimeric and tetrameric structures.

Each seed structure was subjected to replica exchange molecular dynamics (Sugita and Okamoto 1999) (16 replicas, each replica run for 20ns). Structures from the 298K window were output every picosecond and added to the structural library. In total, the structural library contained 60,000 structures (monomers, trimers and tetramers). All of these structures were then clustered using a crude pruning algorithm to ensure that the final set of structures largely retained the structural heterogeneity present in the original 60,000. The final set of structures, including monomers, trimers and tetramers, contained 533 conformers.

We note that each of the replica exchange simulations began with a predominantly helical seed structure because several studies suggest that α-synuclein multimers had significant helical content (Bartels, Choi et al. 2011, Wang, Perovic et al. 2011, Trexler and Rhoades 2012). However, many of the helical multimers rearranged to form strand-rich conformers during the course of the simulations. Hence the final set of 533 structures constitutes a heterogeneous set of conformers that have a range of both helical and strand content.

The final step in our ensemble construction procedure was to assign population weights to each of the 533 structures. One approach to accomplish this is to obtain a single set of weights, $\vec{w} = \{w_i\}_{i=1}^{n}$, such that calculated observables from the final ensemble agree with the corresponding experimentally determined values. However, as we have previously shown, agreement with experiment alone is insufficient to ensure that the constructed ensemble is correct (Fisher, Huang et al. 2010, Fisher and Stultz 2011). This is because the construction of ensembles for disordered systems is an inherently degenerate problem; i.e., the number of experimental constraints pales in comparison to the number of degrees of freedom for the system. To overcome this limitation, we used a previously developed formalism, grounded in Bayesian statistics, to compute the population weights. This Bayesian Weighting (BW) algorithm computes the full posterior distribution over all possible ways of weighting structures in the structural library. From this posterior distribution we can compute an uncertainty measure,

$0 \leq \sigma_{\vec{w}^B} \leq 1$, which describes the spread of the posterior distribution – a metric that is akin to the standard deviation of a Gaussian distribution (Fisher, Huang et al. 2010, Fisher 2012). Our prior work suggests that the numeric value of $\sigma_{\vec{w}^B}$ is correlated with model correctness. When $\sigma_{\vec{w}^B} = 0$, we can be relatively certain that the model is correct. By contrast when $\sigma_{\vec{w}^B} = 1$, it is likely that the ensemble is far from the truth. Nevertheless, when $\sigma_{\vec{w}^B} \neq 0$, we can construct rigorous confidence intervals for quantities of interest that are calculated from the ensemble. The ability to calculate rigorous confidence intervals enables us to perform rigorous hypothesis tests and therefore determine what conclusions we can make from the ensemble with statistical significance.

The final Bayes' ensemble consists of a set of weights, $\vec{w}^B = \{w_i^B\}$, which corresponds to the expected value of the weights calculated from the posterior distribution, and the structural library $S = \{\vec{s}_i\}_{i=1}^n$. The algorithm also ensures that we restrict our analysis to the most important conformers. More precisely, i^{th} structure is excluded from the ensemble when we can say with 95% confidence that $w_i \leq c$. In the end, a total of 311 structures survived this criterion. While the resulting Bayes' ensemble achieves a good fit to the NMR experimental data (Figure IV.1), the corresponding uncertainty parameter is non-zero: $\sigma_{\vec{w}^B} = 0.47$. Consequently, we express ensemble average values along with their corresponding 95% confidence intervals.

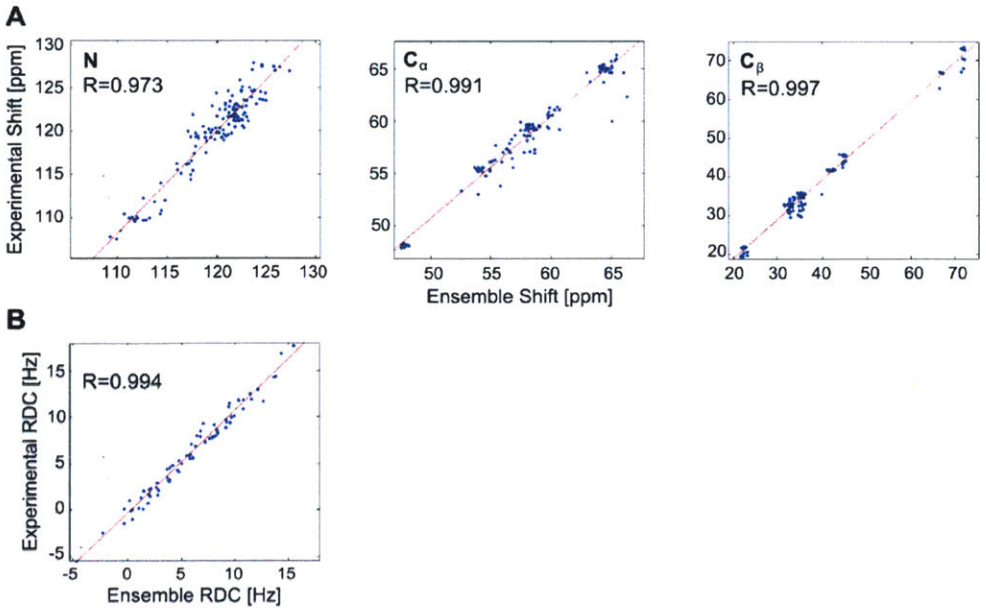

Figure IV.1: Calculated ensemble averages vs. experimental measurements. (A) N, Cα and Cβ chemical shits; (B) N-H residual dipolar couplings. Correlation coefficients for each plot are explicitly shown.

The ensemble is composed mostly of monomeric species ($64.1\% \pm 6.4\%$) with tetrameric species making up the next most common species ($28.2\% \pm 6\%$), and trimeric structures making up only $7.7\% \pm 3.6\%$. Since we have already reported on the types of structures that are sampled in the monomeric protein (Ullman, Fisher et al. 2011), here we focus on the types of multimeric structures that appear in the ensemble. Both trimeric and tetrameric structures mainly come in two forms, either predominantly helical, or predominately strand. A small fraction of multimeric structures contain so little secondary structure that they fall into neither category. Representative structures from each species are shown in Figure IV.2.

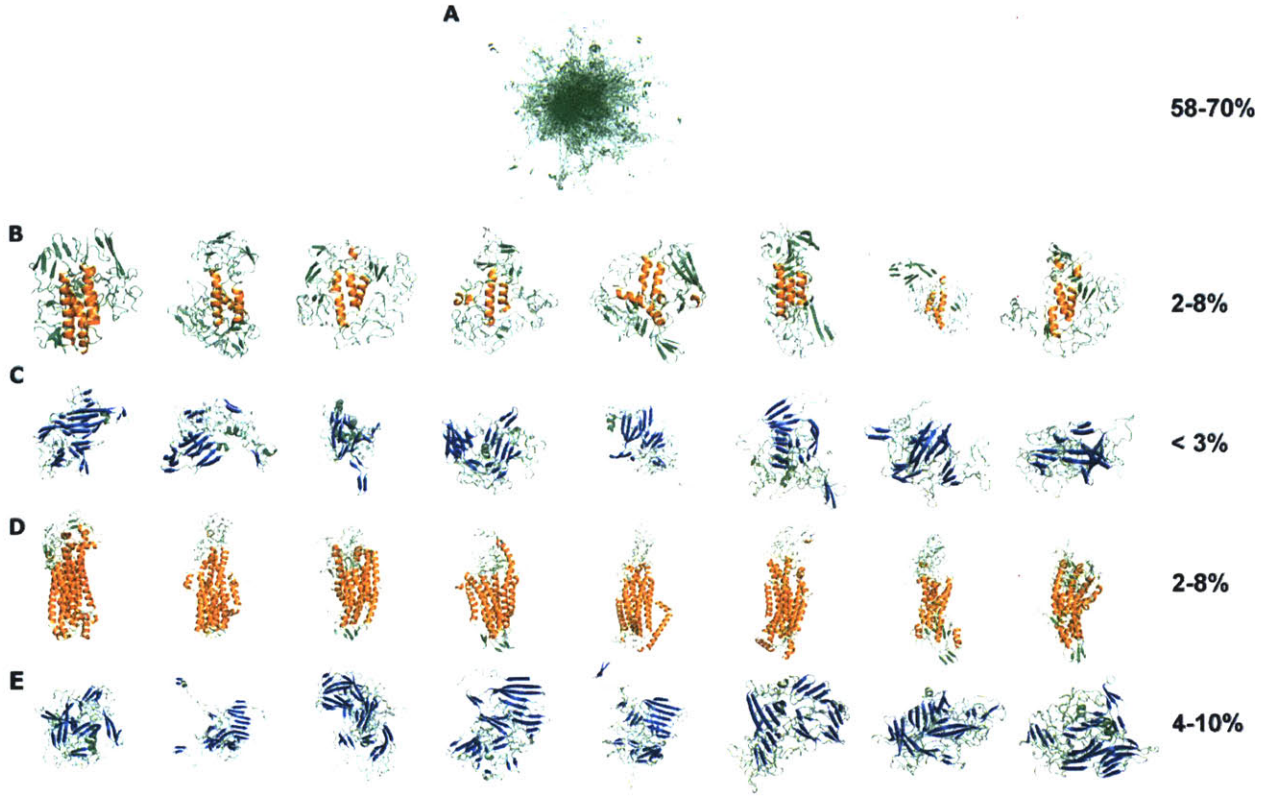

Figure IV.2: Types of α-synuclein structures in our ensemble. Monomers are aligned to each other (A) to demonstrate that they form a structurally heterogeneous set. For the multimeric species, the top 8 structures from each category in terms of secondary structure content are shown: (B) helical-rich trimers; (C) strand-rich trimers; (D) helical-rich tetramers; and (E) strand-rich tetramers.

To determine how each of these multimers may influence α-synuclein self-association, we focus on the position and conformation of the subsequence NAC(8-18), which corresponds to the minimal segment of α-synuclein that can initiate the formation of toxic β-strand rich

aggregates *in vitro* (El-Agnaf and Irvine 2002). This is of particular interest because toxic soluble oligomers of α-synuclein and other related IDPs contain significant β-structure (Volles, Lee et al. 2001, Laganowsky, Liu et al. 2012). Of all the multimeric species in the ensemble, the normalized solvent accessibility of the NAC(8-18) region in helical tetramers is significantly lower than for other types of structures, with an expected value of only 30.6% ± 1.0% (Figure IV.3). For comparison, the solvent exposure of the NAC(8-18) region in the monomeric fraction is 58.6% ± 4.2%. Consequently, helical tetrameric species bury the NAC(8-18) segment relative to the monomeric state. Our findings are consistent with a model where the NAC(8-18) segment initiates the formation of β-rich structures, which then progress to form higher order aggregates. In the β-rich conformers, the NAC(8-18) segment has already been incorporated into β sheet and therefore it is not surprising that their solvent accessibility is reduced. In the helical tetramer the NAC(8-18) segment is hidden in a non-amyloidogenic conformation and is therefore not available to initiate the formation of β-strand rich multimers.

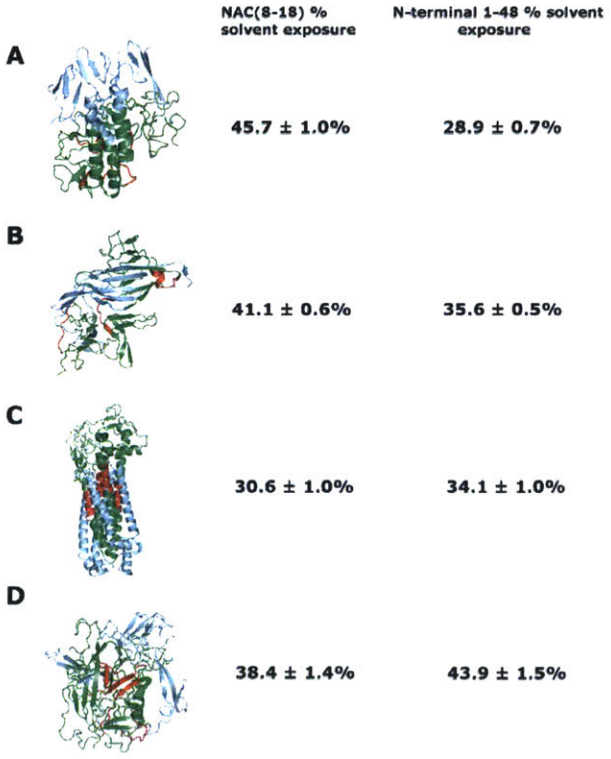

Figure IV.3: Normalized solvent accessibility (± 95% confidence intervals) for the NAC(8-18) region and N-terminal residues 1-48 for (A) helical-rich trimers, (B) strand-rich trimers, (C) helical-rich tetramers and (D) strand-rich tetramers. Representative structures are shown on the left. The N-terminal residues are shown in cyan, the NAC(8-18) in red and the remaining residues in green.

Several studies also suggest that the N-terminal region of α-synuclein may act as an initiation site for the formation of strand-rich oligomeric aggregates. The observation that aggregation-inhibiting small molecules bind preferentially to the N-terminal region of human α-synuclein is consistent with this notion (Cho, Nodet et al. 2009). More importantly, ^{15}N relaxation experiments performed on monomeric mouse α-synuclein (which has faster aggregation kinetics than the human homolog) suggest that the N-terminal region of the protein has decreased backbone flexibility as compared to both a random coil model as well as measurements on human α-synuclein – a finding suggesting that secondary structure formation is more prevalent in the mouse form of the protein (Wu, Kim et al. 2008). It has further been proposed that KTK(E/Q)GV, which are mainly found within the first 48 residues of the protein, can serve as initiation sites for aggregation in mouse α-synuclein (Wu, Kim et al. 2008). Therefore, we computed the average solvent accessibility of the N-terminal 48 residues in each multimeric state to explore the conformation of the N-terminal region of α-synuclein in each of these multimeric states, as shown in Figure IV.3. Helical trimers and tetramers preferentially place the N-terminal region of α-synuclein in positions that are hidden from solvent; i.e., the solvent exposure of these regions is $28.9\% \pm 0.7\%$ and $34.1\% \pm 1.0\%$ for helical trimers and tetramers, respectively. We note that several studies suggest that the N-terminal region of α-synuclein plays a critical role in the formation of helical structures (Bodner, Dobson et al. 2009, Vamvaca, Volles et al. 2009, Bartels, Ahlstrom et al. 2010), hence this region may be important for assembly of the helical tetramer. By contrast, the solvent exposure for the monomeric state is $52.5\% \pm 3.6\%$. Figure IV.4 shows two structures that involve the N-terminal residues in β-sheet formation, highlighting the β-strand propensity of these residues.

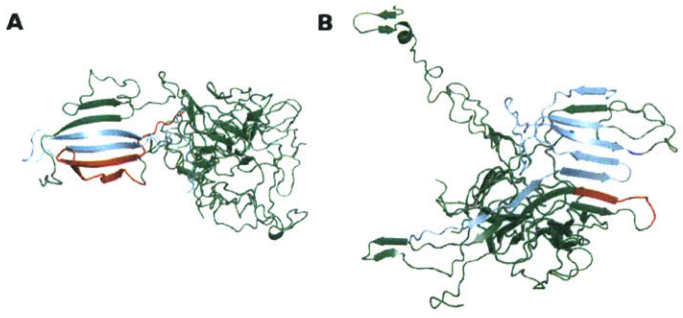

Figure IV.4: Two representative structures of strand-rich tetramers. The N-terminal residues 1-48 of the monomers participating in sheets are shown in cyan. NAC 8-18 residues participating in sheets are shown in red.

Interestingly, however, β-strand rich trimers and tetramers, preferentially have the N-terminal residues 1-48 involved in a sheet that contains the NAC(8-18) segment; i.e., the segment that can initiate α-synuclein aggregation *in vitro* (Figure IV.4). Although it is not clear whether the NAC component or the N-terminal region provides the primary impetus behind the oligomerization propensity of α-synuclein, our data are consistent with a model whereby the initial stages in toxic oligomer formation is the formation of an N-terminal rich β-strand region that contains the NAC(8-18) segment. In this regard, it is interesting that the helical tetrameric species sequesters both of these regions from the surrounding solvent by involving them in the formation of helices, as shown in Figure IV.3, supporting the notion that this structure acts as a non-toxic storage mechanism.

IV.D Conclusions

In this study we constructed an ensemble for the multimeric state of α-synuclein. Our data reveal a number of important insights into the types of structures that multimeric forms of the protein can adopt. Given that generating a comprehensive list of the thermally accessible states of both the monomeric and multimeric protein is not tractable, our goal was to generate a low-resolution description of the dominant states that are available to the protein. However, even with this proviso additional assumptions are needed to make the calculations feasible. In this regard we restricted our sampling of multimeric states to trimers and tetramers; i.e., the primary multimeric states that have been observed when α-synuclein constructs are isolated from *E. coli*, red blood cells and human neuroblastoma cell lines (Bartels, Choi et al. 2011, Wang, Perovic et al. 2011). Replica exchange molecular dynamics (REMD) simulations were used to generate a representative set of heterogeneous set of energetically favorable conformers that served as the template from which a structural ensemble could be built. Given that earlier studies had described the existence of helical trimers and tetramers of α-synuclein, the REMD simulations began using a predefined set of seed structures that were intended to capture conformations that were observed in earlier experiments on α-synuclein multimers. Given that our previous study suggested that the monomeric α-synuclein can sample amphipathic helices,

we generated a model for helical trimers and tetramers assuming that multimeric structures were formed from self-association of these amphipathic helices. A second model seed structure was derived from limited NMR data on α-synuclein at high concentrations. Given the limited number of NOEs obtained, it was not possible to uniquely determine the structure of any tetrameric state; therefore the resulting seed structure serves as fodder for additional simulations, rather than a detailed high-resolution structure of the tetrameric state. Although the REMD simulations began with these seed structures, the resulting trajectories sample a wide region of conformational space leading to the generation of some structures that are very different from the initial seeds (Figures A1 and A2 in the Appendix). The Bayesian Weighting (BW) method is then used to construct a probability density over all possible ways of assigning population weights to structures arising from the trajectories (Fisher, Huang et al. 2010). These data are then used to calculate ensemble average properties with their corresponding confidence intervals.

Given that construction of an ensemble for an intrinsically disordered protein is an inherently degenerate problem, it is important to provide estimates of one's uncertainty in the resulting ensemble (Fisher, Huang et al. 2010, Fisher and Stultz 2011). One advantage of the BW formalism is that it has a built in measure of uncertainty, $0 \leq \sigma_{\vec{w}^B} \leq 1$, that is correlated with model correctness (Fisher, Huang et al. 2010). When $\sigma_{\vec{w}^B} = 0$, we can be relatively certain that the model is correct. By contrast when $\sigma_{\vec{w}^B} = 1$, it is likely that the ensemble is far from the truth. In the present case, this uncertainty parameter is non-zero: $\sigma_{\vec{w}^B} = 0.47$. However, even when the uncertainty parameter is non-zero, one can still quantify the uncertainty in calculated ensemble average quantities via the use of confidence intervals. In this work, we present ensemble averages +/- 95% confidence intervals. Confidence intervals comprise a standard statistical method to quantify uncertainty in an underlying model. The meaning of the confidence interval for the ensemble average $\langle M \rangle$, is that if one calculated $\langle M \rangle$ from many different ensembles (that also fit the experimental data), then those values would fall within the 95% confidence intervals approximately 95% of the time. The 95% confidence interval therefore provides a quantitative measure for the range of values one would see if they constructed many different ensembles. Overall we find that helical tetramers represent a relatively small fraction ($5.1\% \pm 2.9\%$) of an otherwise predominantly disordered, monomeric, ensemble. These findings are consistent with recent bacterial in-cell experiments that suggest

that α-synuclein is predominantly disordered within the crowded intracellular environment (Binolfi, Theillet et al. 2012).

Our data suggest that the multimeric ensemble contains tetrameric states that have significant helical content. However, while some groups have been able to isolate helical tetramers by using gentle purification protocols, the isolation of such structures by other groups has remained elusive (Bartels, Choi et al. 2011, Fauvet, Kamdem et al. 2012, Kang, Moriarty et al. 2012). These latter experiments have led some to conclude that α-synuclein predominantly exists as a disordered monomer under physiologic conditions (Fauvet, Kamdem et al. 2012). We believe our data help to reconcile these seemingly contradictory observations. Our findings argue that helical tetramers are present within the unfolded ensemble, albeit at very low concentrations. Successful isolation of helical tetramers would therefore require additional measures to increase the relative population weight of these states. Indeed, it has been shown that the tetrameric species elute from purification columns in a concentration-dependent manner when the protein is acetylated at its N-terminus (Trexler and Rhoades 2012). This suggests that the relative abundance of this species is a function, in part, of the post-translational state of the protein, the purification protocol, and the protein concentration. These observations are consistent with the notion that the helical tetramer provides a mechanism for *in cellulo* α-synuclein storage when the protein concentration is high. Formation of aggregation resistant helical tetramers may provide a method to sequester non-membrane bound monomers in a form that both prevents them from aggregating and preserves them in a conformation amenable to lipid binding upon dissociation. It is likely that other factors, such as the ionic strength of the medium and presence of divalent metal cations (Dudzik, Walter et al. 2011), would affect the relative stabilities of these various conformations: it has been shown, for instance, that the abundance of β-rich monomers structures increases in the presence of high ionic strength, as well as upon inclusion of Cu2+ (Sandal, Valle et al. 2008).

To understand why helical states are aggregation resistant, we focus on the minimal segment, NAC(8-18), needed to initiate α-synuclein aggregation *in vitro (El-Agnaf and Irvine 2002).* Of all the multimeric states in our ensemble, the solvent exposure of the NAC(8-18) is the lowest for the helical tetramer. Burying the NAC(8-18) segment ensures that is not available to initiate the formation of β-strand rich oligomers. In the β-rich tetramer conformers, the NAC(8-18) segment has already been subsumed in a central β sheet and therefore it is not

surprising that its solvent accessibility is reduced relative to the monomeric state. Our findings are consistent with a model where the NAC(8-18) segment initiates the formation of β-rich tetramer structures, which then progress to form higher order aggregates.

The appearance of strand-rich states in our ensemble is somewhat surprising given that previously published CD spectra of multimeric α-synuclein suggested that the protein had considerable helical content on average (Bartels, Choi et al. 2011, Wang, Perovic et al. 2011). Although the reported CD spectra have distinct minima at 208nm and 222nm – a finding indicative of considerable helical content – estimating the precise helical content from CD spectra alone is problematic (Manavalan and Johnson 1985, Greenfield 2007). For example, we used several different algorithms to quantify the helical content from the published CD spectrum of α-synuclein isolated from human red blood cells (Bartels, Choi et al. 2011), and depending on the algorithm used, the amount of helix varied from 10% to 80%. Hence, while the CD spectrum suggests that the helical content of the tetrameric species is higher than that of the monomeric protein, quantifying the amount of helicity from the CD spectrum alone is a non-trivial exercise. In addition, the multimeric ensemble was generated using data from NMR experiments that were performed at a concentration (0.5mM) that was at least an order of magnitude greater than the concentration used for the CD experiments (~0.02mM). This is important because the concentration of α-synuclein *in vitro* can influence its secondary structure propensity and the precise effect may vary on the post-translational state of the protein (Jarrett and Lansbury 1993, Iwai, Yoshimoto et al. 1995, Trexler and Rhoades 2012). Therefore it is not clear whether the published CD spectrum reflects the structure of α-synuclein under the conditions used for the NMR experiments.

Lastly, we note that a limitation of our study is that the NMR data were obtained on an α-synuclein construct that contains a 10-residue N-terminal extension relative to the wild-type protein. While the experimental data provided useful constraints that could be fruitfully applied to generate an ensemble, α-synuclein isolated from human neuroblastoma and red blood cell lines does not have an N-terminal extension and instead is acetylated at the N-terminus (Bartels, Choi et al. 2011). Nevertheless, our construct shares important characteristics with the N-acetylated protein. First, the monomeric form of the construct bearing a 10-residue N-terminal extension has a CD spectrum that is similar to that of the monomeric N-terminal acetylated form of α-synuclein (Fauvet, Kamdem et al. 2012) and both constructs form tetrameric structures with

increased α-helical content (Bartels, Choi et al. 2011, Wang, Perovic et al. 2011, Trexler and Rhoades 2012). Lastly, monomeric forms of both constructs have similar aggregation profiles while the tetrameric forms of both constructs do not aggregate (Bartels, Choi et al. 2011, Wang, Perovic et al. 2011). These similarities suggest that acetylation of the N-terminal and the 10 residues elongation of the N terminal region in α-synuclein serve a similar purpose with regard to their effect on the α-synuclein, albeit N-terminal acetylation may result in more dramatic effects to the conformational distribution of the protein relative to the N-terminal extension. Nonetheless, since the sequence of this construct differs slightly from the wild-type protein, we cannot exclude the possibility that wild-type α-synuclein isolated from other cell types, such as neurons or red blood cells, may not be well described by the ensemble presented here.

IV.E Materials and Methods

IV.E.1 Generation of seed structures

Our previous study on α-synuclein suggested that the monomeric, protein can sample amphipathic helices, which could in principle self-associate to form helical trimers and tetramers (Ullman, Fisher et al. 2011).

All simulations used a model of α-synuclein that did not include the 10-residue N-terminal extension. An initial trimeric structure of the protein was generated by taking a monomer from the monomeric α-synuclein ensemble that has an amphipathic helix between residues 52 and 64 and threading the helix to a three-helix bundle from a crystal structure of myosin (PDB ID code 3GN4) (Mukherjea, Llinas et al. 2009), where the hydrophobic faces of the amphipathic helix were oriented such that they face inwards. An initial tetrameric structure was generated by threading the same monomer to a four-helix bundle from a crystal structure of ferritin (PDB ID code 1FHA) (Lawson, Artymiuk et al. 1991, Berman, Westbrook et al. 2000). These structures were chosen from the PDB such that the helix bundles in the structure used for threading the monomer were of sufficient length to accommodate the entire 12-residue helix in our monomer structure, while retaining a high enough resolution to be informative. A second initial helical tetrameric model was constructed using the available NMR data (Wang, Perovic et

al. 2011). The model derived from the NMR data was obtained from a limited set of NOEs; i.e., we were not able to identify a sufficient number of sequential (Hα-HN *i, i*+3) NOEs in ^{15}N-edited NOESY spectra (see below). Consequently, the resulting model is not intended to represent a "high-resolution" structure of the helical tetramer. Instead, its only purpose is to serve as a structure (derived from limited experimental data) that is the starting point for additional simulations. More generally, each seed structure serves as a starting point from which to begin more extensive sampling.

IV.E.2 Generation of α-synuclein structural library

The conformational space of α-synuclein was sampled by subjecting the initial seed structures to replica exchange molecular dynamics (REMD) simulations (Sugita and Okamoto 1999). Each initial structure underwent REMD with the EEF1 (Lazaridis and Karplus 1999) implicit solvent model as implemented in the CHARMM (Brooks, Bruccoleri et al. 1983) force field. Sixteen replicas were used, with temperatures equally spaced in 5K increments over the 293-368K range. Prior studies of IDPs with this implicit solvent model have yielded useful insights (Huang and Stultz 2008, Fisher, Huang et al. 2010, Ullman, Fisher et al. 2011). Initially, higher temperature replicas were explored, along with quenched molecular dynamics simulations at higher temperatures, but we found that these led to dissociation of multimers into monomers free of intermolecular contacts. We therefore limited the highest temperature to 368K, the highest temperature at which intermolecular contacts were retained in oligomers for the duration of the trajectory. Each replica was run for 20 ns, and structures were collected at each picosecond. A total of 20,000 conformations per REMD simulation were collected, all from the 298K window, making a total of 60,000 conformations for the trimeric and tetrameric structures.

The set of 60,000 structures was pruned down by enforcing a minimum pairwise RMSD of 9Å to ensure that the resulting library would span a range of heterogeneous conformations. The resulting set contained 234 structures. These were then combined with 299 monomer structures from a previously constructed monomeric ensemble of α-synuclein (Ullman, Fisher et al. 2011) to yield our structural library $S = \{\vec{s}_i\}_{i=1}^{533}$ of 533 conformers.

IV.E.3 Generation of the ensemble and calculation of confidence intervals

To obtain the set of weights associated with each conformer in our structural library, we employ the Variational Bayesian Weighting algorithm (VBW) previously described (Fisher 2012), which is a variational approximation to a Bayesian Weighting formalism used in the past (Fisher, Huang et al. 2010, Ullman, Fisher et al. 2011). This algorithm generates a posterior distribution $f_{\vec{W}|\vec{M},S}(\vec{w}\,|\,\vec{m},S)$ for the weights, conditioned on the set of 533 structures, and the provided experimental measurements. The form of the posterior distribution is dictated by Bayes' rule:

$$f_{\vec{W}|\vec{M},S}(\vec{w}\,|\,\vec{m},S)=\frac{f_{\vec{M}|\vec{W},S}(\vec{m}\,|\,\vec{w},S)f_{\vec{W}|S}(\vec{w}\,|\,S)}{f_{\vec{M}|S}(\vec{m}\,|\,S)} \tag{1}$$

where the term $f_{\vec{W}|S}(\vec{w}\,|\,S)$ is the prior distribution and $f_{\vec{M}|\vec{W},S}(\vec{m}\,|\,\vec{w},S)$ is the likelihood function for the experimental observations $\vec{m}$, whose full descriptions can be found in the original publication of the method (Fisher 2012). Experimental observables, specifically Cα, Cβ, N, H and Hα chemical shifts from a previous work (Wang, Perovic et al. 2011) in combination with backbone NH residual dipolar couplings (RDCs), were used. Predicted measurements for each conformer were generated using SHIFTX (Neal, Nip et al. 2003) for chemical shifts and PALES (Zweckstetter 2008) for residual dipolar couplings. Residual dipolar couplings were uniformly scaled to account for uncertainty in the magnitude of the alignment tensor. Similarly, like-atom chemical shifts were uniformly offset to account for uncertainty in chemical shift referencing. To increase computational efficiency and analytical tractability, an approximation from variational Bayesian inference was applied by choosing a simpler probability density function (PDF) (Fisher 2012), which approximates the full posterior distribution, calculated from equation (1). For a vector of weights, a natural choice is the Dirichlet distribution with parameters $\{\alpha_i>0\}_{i=1}^{N}$. This results in an approximate PDF for the weights (Fisher 2012):

$$g(\vec{w}\,|\,\vec{\alpha},S)=\frac{\Gamma(\alpha_0)}{\sum_{i=1}^{n}\Gamma(\alpha_i)}\prod_{i=1}^{N}w_i^{\alpha_i-1} \tag{2}$$

where α_i is the Dirichlet parameter associated with weight i and $\alpha_0 = \sum_i \alpha_i$. The Kullback-Leibler distance (i.e., the KL divergence) between $g(\vec{w}\,|\,\vec{\alpha},S)$ and $f_{\vec{W}|\vec{M},S}(\vec{w}\,|\,\vec{m},S)$ is then minimized to find the optimal set of Dirichlet parameters, $\vec{\alpha}' = \{\alpha_i'\}_{i=1}^{N}$, which provides an approximation to the true posterior from which one can easily calculate quantities of interest.

We then compute the Bayes estimate for the weights $\vec{w}^B = \{w_i^B\}$, which is the expected value of the vector of weights over the new approximate posterior distribution:

$$\vec{w}^B = \int d\vec{w} g(\vec{w}\,|\,\vec{\alpha}',S)\vec{w} \tag{3}$$

The Bayes estimate can be calculated from the Dirichlet PDF according to:

$$w_i^B = \frac{\alpha_i'}{\alpha_0'} \tag{4}$$

where $\alpha_0' = \sum_i \alpha_i'$. The uncertainty parameter $\sigma_{\vec{w}^B}$, called the posterior expected divergence, corresponds to the average distance from the Bayes weights over the entire space of weights:

$$\sigma_{\vec{w}^B} = \sqrt{\int d\vec{w}\Omega^2(\vec{w}^B,\vec{w})g(\vec{w}\,|\,\hat{\alpha}',S)}\ \vec{w}^B \tag{5}$$

where $\Omega^2(\vec{w}^B,\vec{w})$ is the Jensen-Shannon divergence, a metric which quantifies the distance between the vectors $\vec{w}^B$ and $\vec{w}$ (Fisher, Huang et al. 2010).

The covariance between the weights of conformers i and j can be calculated analytically from:

$$\mathrm{cov}(w_i,w_j) = \frac{\alpha_i'\alpha_0'\delta_{ij} - \alpha_i'\alpha_j'}{\alpha_0'^2(\alpha_0'+1)} \tag{6}$$

where δ_{ij} is the Kronecker Delta function. Any quantity D that can be calculated for a given conformer can then be assigned a variance across the ensemble according to:

$$\mathrm{var}(D) = \sum_i \sum_j D_i D_j \,\mathrm{cov}(w_i,w_j) \tag{7}$$

95% confidence intervals can then be computed using a Gaussian approximation from $CI = 1.54 \times 1.96 \times \sqrt{\mathrm{var}(D)}$, where 1.54 is an empirical factor relating the variational approximation of the posterior distribution to the true posterior distribution under the complete BW formalism (Fisher 2012).

A backward elimination procedure starting with our initial structural library of 533 conformers was used to ensure that the ensemble only contained essential structures. The procedure computed the VBW posterior distribution iteratively. After each iteration, all non-essential structures were identified by finding the largest set I such that the joint probability that each weight of the structures in I fell below a cut-off exceeded a chosen confidence level, i.e. $\prod_{i \in I} P(w_i \leq c) \geq 1 - \theta$ where $P(\cdot)$ denotes the cumulative distribution function of the weights. The cut-off (c) and confidence level (θ) were set to 0.005 and 0.05 (95%), respectively. Each of the non-essential structures in I were removed and the weighting procedure repeated. This process was iterated until convergence, i.e. until the cardinality of I was zero.

IV.E.4 Secondary structure assignments

Secondary structure was assigned using DSSP (Kabsch and Sander 1983). A residue was assigned to the class of 'helix' if it was assigned as α-helix, π-helix or 3-10 helix by DSSP. Similarly, a residue was assigned to the class of 'strand' if it was assigned as a bridge or extended by DSSP. The remaining assignments were grouped into the class of 'other'. Structures appearing in the uppermost quartile of tetramers ranked by helical content were classified as helical tetramers, and structures in the uppermost quartile of tetramers ranked by strand content were classified as strand tetramers. Trimers were classified in the same manner.

IV.E.5 Solvent accessibility calculations

Solvent accessible surface area (SASA) was calculated for each conformation using CHARMM(Brooks, Bruccoleri et al. 1983). Since only the backbone atoms N, H, C, Cα and O are involved in the formation of secondary structure, only SASA values for these atoms were considered. The solvent accessibility for the entire protein was computed by summing each atom's SASA value and normalized by dividing the result by the SASA of the α-synuclein backbone atoms when in a fully extended conformation.

IV.E.6 NMR studies

It is important to note that these NMR studies were insufficient to uniquely determine the structure of a helical tetrameric state (primarily due to an insufficient number of measured NOEs). Hence, the structure arising from these studies represents a model that only serves as the starting point for further simulations, as opposed to a well-defined structure for the helical tetramer.

Samples of ^{15}N and ^{13}C labeled αSyn for NMR spectroscopy were prepared using uniformly ^{13}C- and ^{15}N-labeled media (supplemented M9 media, ^{13}C source being glucose). NMR samples were typically prepared to a final concentration of ~0.5 mM in 100 mM Tris•HCl pH 7.4, 100 mM NaCl, 0.1% BOG, 10% glycerol, 10% D_2O. All NMR spectroscopy was performed on a Bruker Avance 800 NMR spectrometer operating at 800.13 MHz (^1H), 81.08 MHz (^{15}N) and 201.19 MHz (^{13}C) and equipped with a TCI cryoprobe and pulsed field gradients. Experiments used for sequential resonance assignments include three-dimensional (3D) experiments HNCA, HNCACB, ^{15}N-HSQC TOCSY and ^{15}N-HSQC NOESY. Quadrature detection was obtained in the ^{15}N dimension of 3D experiments using sensitivity-enhanced gradient coherence selection(Kay, Keifer et al. 1992), and in the ^{13}C dimension using States-TPPI, with coherence selection obtained by phase cycling. In all cases, spectral widths of 8802.82 Hz (^1H) and 2920.56 Hz (^{15}N) were used. For ^{13}C, spectral widths of 6451.61 Hz (HNCA) and 15105.74 Hz (HNCACB) were used. All experiments were performed at 298 K unless otherwise noted. NMR data were processed using TOPSPIN (Bruker Biospin Inc.), and data analyzed using either TOPSPIN or SPARKY (Goddard and Kneller).

^1H-^{15}N, $^{13}C'$-^{15}N and $^{13}C'$-$^{13}C\alpha$ residual dipolar couplings (RDCs) were recorded for a ^{15}N- and ^{13}C-labeled wild-type αSyn oligomer sample in the presence and absence of alignment media using a standard IPAP-HSQC sequence or a variation of a standard HNCO pulse sequence. Sample alignment was accomplished using a 5% polyacrylamide stretched gel. We chose to use PA rather than bicelle or liquid crystalline phases for alignment because such phases contain long chain hydrocarbon moieties that might be expected to bind αSyn and could interfere with oligomer formation.

The stretched gel was prepared using a commercial apparatus (New Era, Vineland, NJ) according to the manufacturer's protocol and following guidelines by A. Bax.(Bax 2003) After polymerization was complete, the gel was dialyzed against water overnight at room temperature,

and then incubated with a 0.5 mM αSyn sample in standard NMR buffer for 48 h at 4 °C. The diameter of the gel was 6.0 mm before and 4.2 mm after stretching. Alignment was confirmed by observing the residual quadrupolar splitting (9.4 Hz) of the ^{2}H water signal.

We used solution NMR to localize the transient formation of α-helices in αSyn. Resonance assignments were made using standard methods (HNCO, HN(CO)CA, HNCA, HNCACB, ^{15}N-edited NOESY and TOCSY). Although a high degree of spectral overlap is present even in three-dimensional data sets, we were able to identify a number of sequential (Hα-HN *i, i*+3) NOEs in ^{15}N-edited NOESY spectra to confirm the transient existence of α-helical structure between residues Phe4-Thr43 and His50-Asn103. In many cases, these NOEs are quite weak, consistent with fractional occupancy, however, only the most reliable (strongest) experimental NOEs were used in model construction. Note that if long stretches of NOEs interrupted by several residue pairs without NOEs were observed, the missing pairs were included in the helical restraints applied in XPLOR-NIH. A total of 73 unique inter-residue NOEs per monomer were used to construct a model for the helical tetramer.

Given the relatively small number of NOEs any structure arising from these data merely represents a model (derived from limited experimental data) that serves as fodder for additional simulations, rather than a detailed high-resolution structure of the tetrameric state.

A combined torsional and Cartesian dynamics simulated annealing method was used to calculate an average tetramer structure using XPLOR-NIH v. 2.18(Schwieters, Kuszewski et al. 2003). Secondary structural restraints were applied to those regions of the polypeptide identified as forming α-helical structure from sequential NOEs. RDC restraints were applied for residues 1-103 and in some cases, non-crystallographic symmetry restraints were applied to residues 4-36, 47-85 and 89-98. Preliminary structures were crafted manually using PyMOL (Schrodinger 2010), and initial values for alignment tensors determined by singular value decomposition (SVD) using the program PALES (Zweckstetter 2008). As refinement proceeded, best-fit structures were used to recalculate the alignment tensors via a combined SVD-least squares fit which permits the rhombic terms to be fixed at zero. This was applied iteratively until no further improvements of fit were observed. PyMOL was also used for visualization of the structures generated by XPLOR-NIH. Proton chemical shifts were referenced directly to the water signal at 4.7 ppm, while ^{15}N and ^{13}C shifts were indirectly referenced (Wishart, Bigam et al. 1995). All

NMR experiments were performed by Iva Perovic and Thomas Pochapsky. Structural models for the multimeric state of α-synuclein will be freely available via http://www.rle.mit.edu/cbg.

Chapter V

Conclusions

The leitmotif of this work is that IDPs in solution can be modeled as a finite set of energetically favorable structures, where each structure corresponds to an energy minimum over a complex energy landscape. We introduced the notion of IDPs as a class of proteins that lack a well-defined structure in solution. IDPs can adopt a range of rapidly interconverting dissimilar conformations. As such, the energy landscape of an IDP can be described by a complex relatively "flat" surface where energy minima are separated by very small barriers.

In practice it is useful to model IDPs using sets of conformations that represent, at low resolution, the types of structures that the protein can adopt during its biological lifetime. By using a finite set of conformations we essentially specify the resolution of our model, the smaller the number of conformations used, the lower (i.e., poorer) the model resolution.

To build ensembles for IDPs we used a previously developed method that is grounded in Bayesian statistics. The advantage of this method is that it overcomes a longstanding, but rarely acknowledged, shortcoming of the IDP construction process – namely that building ensembles that agree with experiment is an inherently degenerate problem. More precisely, the number of experimental observables used to build the model pales in comparison with the number of degrees of freedom needed to determine the contribution of each structure to the model. Put in other words, for a given set of conformations, one can reproduce the available experimental data using many different ensembles. The Bayesian framework provides a statistically sound platform for generating ensembles. It has the added benefit that it enables us to quantify our uncertainty in the resulting structural models. Moreover, with the Bayesian framework we can calculate rigorous confidence intervals for quantities that are calculated from the ensemble. We apply these methods to the IDP, α-synuclein, which plays a role in the pathogenesis of Parkinson's disease. Our goal was to generate a structural ensemble for α-synuclein that agrees with experimental data, and to use this model to better understand the mechanism underlying α-synuclein self-association – the process that has been associated with neuronal death and dysfunction.

The first step in any ensemble generation method is the construction of a diverse set of energetically favorable conformations. We demonstrated that using a segment-based approach efficiently generates a diverse set of structures. We then combined these structures with experimental data to develop an ensemble for α-synuclein. Our data helps explain the proteins ability to adopt a variety of structures under different experimental conditions.

Lastly we developed an ensemble for multimeric α-synuclein. This study is particularly poignant because recent data suggests that α-synuclein exists as a helical tetramer within the intracellular environment, and that this tetrameric state is aggregation resistant. We therefore collaborated with Prof. Thomas Pochapsky (Brandeis University) to generate a model for an α-synuclein construct that forms multimeric states (one of which is a helical multimer) in solution. We found that the sample is in fact primarily monomeric and that only a small fraction is tetrameric. An analysis of the helical tetrameric state argues that the tetramer does not aggregate because it places aggregation prone segments in environments that are hidden from solvent.

While these data are encouraging, much still needs to be done to make these methods generally applicable to other IDPs. One of the major challenges in generating structural models for IDPs, from the computational point of view, is the lack of readily available experimental data. Since experimental data serves as a benchmark for creating a reasonable structural ensemble, the more data one can incorporate in creating the ensemble the better models one can build. Therefore creating a platform similar to the RCSB Protein Data Bank (Berman, Westbrook et al. 2000), solely devoted to disordered proteins, where one can have access to the raw experimental data such as NMR chemical shifts, SAXS, RDCs to name a few, can help in progressing the efforts to create computational structural models for many IDPs. If experimental and computational groups are encouraged to deposit their data, it can help identify knowledge gaps as well as encourage collaborations by generating a community. DisProt database has already made some progress on this matter, by indexing and referencing many disordered proteins and proteins containing disordered regions (Sickmeier, Hamilton et al. 2007). However, it does not provide access to raw data, only reference to articles and the experimental conditions.

In order to assess the agreement with experiments we calculate predicted experimental observables from the structure coordinates and compared those with the actual experimental observable. For example, we used SHIFTX (Neal, Nip et al. 2003) to calculate NMR chemical shifts and PALES (Zweckstetter 2008) to calculate NH RDCs from individual conformers.

Errors associated with the prediction of experimental data from three-dimensional structures add to the uncertainty in the underlying model. In particular we found that the predicted error for Hα chemical shifts, when using SHIFTX, is on par with the chemical shift dispersion for typical IDPs. Using these data therefore does not provide additional information in model construction and we had to discard of it. It is therefore essential to create more accurate algorithms, thus ensuring that one can maximize the number of experimental observations used in building the models.

Appendix A

Calculating the radius of gyration range from its distribution

The distribution of the radius of gyration for a linear chain non-perturbed by volume exclusion is given by(Flory and Fisk 1966):

$$(1)\ P(R_g) = A(R_g{}^2)^3 exp\left(-\frac{7}{2}\left(\frac{R_g{}^2}{\langle R_g{}^2\rangle}\right)\right)$$

Where A is a normalization constant

Using:

$$(2)\ \int_0^\infty x^n exp[-ax^2]dx = \begin{cases} \frac{(2k-1)!!}{2^{k+1}a^k}\sqrt{\frac{\pi}{a}} & n = 2k, k \in \{\mathbb{Z}\}, a > 0 \\ \frac{k!}{2a^{k+1}} & n = 2k+1, k \in \{\mathbb{Z}\}, a > 0 \end{cases}$$

We can calculate the mean radius of gyration and the mean squared radius of gyration:

$$(3)\ \langle R_g\rangle = \int_0^\infty P(R_g)R_g dR_g =$$

$$\int_0^\infty AR_g{}^7 exp\left(-\frac{7}{2}\left(\frac{R_g{}^2}{\langle R_g{}^2\rangle}\right)\right)dR_g =$$

$$A\int_0^\infty R_g{}^7 exp\left(-\frac{7}{2}\left(\frac{R_g{}^2}{\langle R_g{}^2\rangle}\right)\right)dR_g =$$

$$A\frac{3!}{2}\left(\frac{7}{2\langle R_g{}^2\rangle}\right)^{-4} = 3A\left(\frac{2\langle R_g{}^2\rangle}{7}\right)^4$$

$$(4)\ \langle R_g{}^2\rangle = \int_0^\infty P(R_g)R_g{}^2 dR_g =$$

$$\int_0^\infty AR_g{}^8 exp\left(-\frac{7}{2}\left(\frac{R_g{}^2}{\langle R_g{}^2\rangle}\right)\right)dR_g =$$

$$A\int_0^\infty R_g{}^8 exp\left(-\frac{7}{2}\left(\frac{R_g{}^2}{\langle R_g{}^2\rangle}\right)\right)dR_g =$$

$$A\frac{7!!}{2^5}\left(\frac{7}{2}\left(\frac{1}{\langle {R_g}^2\rangle}\right)\right)^{-4}\sqrt{\pi}\left(\frac{7}{2}\left(\frac{1}{\langle {R_g}^2\rangle}\right)\right)^{-0.5}$$

$$= A\left(\frac{2\langle {R_g}^2\rangle}{7}\right)^4\frac{7!!}{2^5}\left(\frac{2\pi}{7}\right)^{0.5}\langle {R_g}^2\rangle^{0.5}$$

Using the expression for the mean radius of gyration from equation (3) we find:

$$(5)\ \langle {R_g}^2\rangle^{0.5} = \frac{\langle R_g\rangle}{3}\frac{7!!}{2^5}\left(\frac{2\pi}{7}\right)^{0.5} = 1.04\langle R_g\rangle$$

We use the Flory power law(Flory 1953):

$$(6)\ \langle R_g\rangle = R_0 N^{\upsilon}$$

where N is the number of monomers in the polymer chain (in our case 140), R_0 is a proportionality constant that is a function of the polymer's mean persistence length among other things, and υ is the exponential scaling factor. We obtained the values of the scaling factor and the constant from an extensive SAXS study of 28 proteins under strong denaturing conditions, where the best-fit values for the power law are: $R_0 = 1.927^{+0.271}_{-0.238}$Å and $\upsilon = 0.598 \pm 0.028$ where the bounds represent 95% confidence intervals (Kohn, Millett et al. 2004, Kohn, Millett et al. 2005). To extend the range for the probability distribution we chose to use the higher bounds for the formula of the mean radius of gyration (i.e. $R_0 = 2.198$Å and $\upsilon = 0.626$). Therefore the calculated average radius of gyration is $\langle R_g\rangle = 48.5$Å.

Sampling from the distribution in equation (1) where we use the relation in equation 7 and the calculated value for $\langle R_g\rangle$, we find that the 95% percentile boundary to be 71Å.

Appendix B

Supporting figures for chapter IV

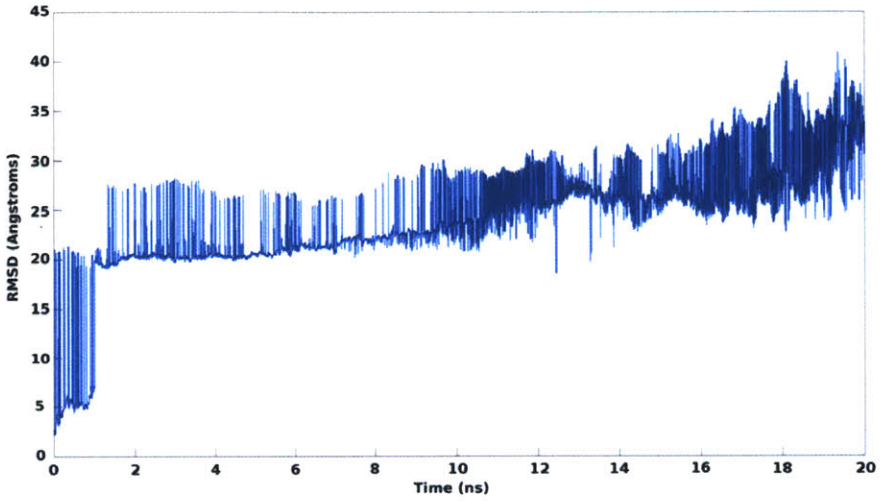

Figure A1: Conformational heterogeneity of a single REMD simulation. Shown is the Cα-RMSD of a structure at time *t* in the 298K temperature replica compared to the original seed structure (in this case the threaded helical tetramer).

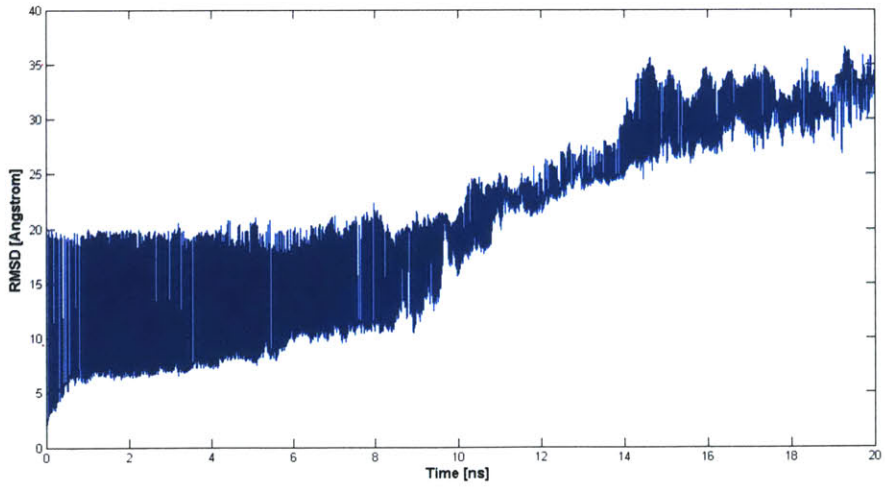

Figure A2: Conformational heterogeneity of a single REMD simulation. Shown is the Cα-RMSD of a structure at time *t* in the 298K temperature replica compared to the original seed structure (in this case the structure derived from limited NMR data).

Appendix C

List of Acronyms

BW	Bayesian Weighting
CD	Circular Dichroism
CHARMM	Chemistry at HARvard Molecular Mechanics
CPU	Central Processing Unit
DLB	Dementia with Lewy bodies
DSSP	Define Secondary Structure of Proteins
EEF1	Effective Energy Function 1
ESR	Electron Spin Resonance
HSQC	Heteronculear Single Quantum Correlation
IDP	Intrinsically Disordered Protein
KL	Kullback-Leibler
MD	Molecular Dynamics
MSA	Multiple System Atrophy
NAC	Non-Amyloid β Component
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Enhancement
NOESY	Nuclear Overhauser Enhancement SpectroscopY
PALES	Prediction of AlignmEnt from Structure
PD	Parkinson's disease
PDB	Protein Data Bank
PDF	Probability Density Function
PHF6	Paired Helical Filaments 6
PRE	Paramagnetic Relaxation Enhancement
RBC	Red Blood Cell
RCSB	the Research Collaboratory for Structural Bioinformatics
RDC	Residual Dipolar Coupling
REMD	Replica Exchange Molecular Dynamics

RMSD	Root-Mean-Square Deviation
RMSE	Root-Mean-Square Error
SASA	Solvent-Accessible Surface Area
SAXS	Small-Angle X-ray Scattering
STRIDE	secondary STRuctural IDEntification
SVD	Singular Value Decomposition
TIP3P	Transferable Intermolecular Potential 3P
TOCSY	TOtal Correlation SpectroscopY
VBW	Variational Bayesian Weighting
WT	Wild-Type

Bibliography

Allison, J. R., P. Varnai, C. M. Dobson and M. Vendruscolo (2009). "Determination of the Free Energy Landscape of α-Synuclein Using Spin Label Nuclear Magnetic Resonance Measurements." Journal of the American Chemical Society **131**(51): 18314-18326.

Andrade, M. A., P. Chacon, J. J. Merelo and F. Moran (1993). "Evaluation of Secondary Structure of Proteins from UV Circular-Dichroism Spectra Using an Unsupervised Learning Neural-Network." Protein Engineering **6**(4): 383-390.

Apetri, M. M., N. C. Maiti, M. G. Zagorski, P. R. Carey and V. E. Anderson (2006). "Secondary Structure of α-Synuclein Oligomers: Characterization by Raman and Atomic Force Microscopy." Journal of molecular biology **355**(1): 63-71.

Appel-Cresswell, S., C. Vilarino-Guell, M. Encarnacion, H. Sherman, I. Yu, B. Shah, D. Weir, C. Thompson, C. Szu-Tu, J. Trinh, J. O. Aasly, A. Rajput, A. H. Rajput, A. Jon Stoessl and M. J. Farrer (2013). "α-Synuclein p.H50Q, a Novel Pathogenic Mutation for Parkinson's Disease." Mov Disord **28**(6): 811-813.

Auluck, P. K., G. Caraveo and S. Lindquist (2010). "α-Synuclein: Membrane Interactions and Toxicity in Parkinson's Disease." Annu Rev Cell Dev Biol **26**: 211-233.

Banner, D. W., M. Kokkinidis and D. Tsernoglou (1987). "Structure of the ColE1 Rop Protein at 1.7 Å Resolution." Journal of Molecular Biology **196**(3): 657-675.

Bartels, T., L. S. Ahlstrom, A. Leftin, F. Kamp, C. Haass, M. F. Brown and K. Beyer (2010). "The N-Terminus of the Intrinsically Disordered Protein α-Synuclein Triggers Membrane Binding and Helix Folding." Biophysical Journal **99**(7): 2116-2124.

Bartels, T., J. G. Choi and D. J. Selkoe (2011). "α-Synuclein Occurs Physiologically as a Helically Folded Tetramer that Resists Aggregation." Nature **477**: 107-110.

Bax, A. (2003). "Weak Alignment Offers New NMR Opportunities to Study Protein Structure and Dynamics." Protein Science **12**(1): 1-16.

Bellucci, A., M. Zaltieri, L. Navarria, J. Grigoletto, C. Missale and P. Spano (2012). "From α-Synuclein to Synaptic Dysfunctions: New Insights into the Pathophysiology of Parkinson's Disease." Brain Research.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne (2000). "The Protein Data Bank." Nucleic Acids Research **28**(1): 235-242.

Bernado, P., C. W. Bertoncini, C. Griesinger, M. Zweckstetter and M. Blackledge (2005). "Defining Long-Range Order and Local Disorder in Native α-Synuclein Using Residual Dipolar Couplings." Journal of the American Chemical Society **127**(51): 17968-17969.

Bernadó, P., L. Blanchard, P. Timmins, D. Marion, R. W. H. Ruigrok and M. Blackledge (2005). "A Structural Model for Unfolded Proteins from Residual Dipolar Couplings and Small-Angle X-Ray Scattering." Proceedings of the National Academy of Sciences of the United States of America **102**(47): 17002.

Bertoncini, C. W., C. O. Fernandez, C. Griesinger, T. M. Jovin and M. Zweckstetter (2005). "Familial Mutants of α-Synuclein with Increased Neurotoxicity Have a Destabilized Conformation." Journal of Biological Chemistry **280**(35): 30649-30652.

Bertoncini, C. W., Y. S. Jung, C. O. Fernandez, W. Hoyer, C. Griesinger, T. M. Jovin and M. Zweckstetter (2005). "Release of Long-Range Tertiary Interactions Potentiates Aggregation of Natively Unstructured α-Synuclein." Proceedings of the National Academy of Sciences of the United States of America **102**(5): 1430.

Binolfi, A., R. M. Rasia, C. W. Bertoncini, M. Ceolin, M. Zweckstetter, C. Griesinger, T. M. Jovin and C. O. Fernandez (2006). "Interaction of α-Synuclein with Divalent Metal Ions Reveals Key Differences: A Link Between Structure, Binding Specificity and Fibrillation Enhancement." Journal of the American Chemical Society **128**(30): 9893-9901.

Binolfi, A., F. X. Theillet and P. Selenko (2012). "Bacterial In-Cell NMR of Human α-Synuclein: a Disordered Monomer by Nature?" Biochemical Society Transactions **40**: 950-U292.

Blanco, F. J., G. Rivas and L. Serrano (1994). "A Short Linear Peptide That Folds into a Native Stable β-Hairpin in Aqueous Solution." Nature Structural Biology **1**(9): 584-590.

Bodner, C. R., C. M. Dobson and A. Bax (2009). "Multiple Tight Phospholipid-Binding Modes of α-Synuclein Revealed by Solution NMR Spectroscopy." Journal of Molecular Biology **390**(4): 775-790.

Boomsma, W., J. Ferkinghoff-Borg and K. Lindorff-Larsen (2014). "Combining Experiments and Simulations Using the Maximum Entropy Principle." Plos Computational Biology **10**(2).

Borbat, P., T. F. Ramlall, J. H. Freed and D. Eliezer (2006). "Inter-Helix Distances in Lysophospholipid Micelle-Bound α-Synuclein from Pulsed ESR Measurements." J. Am. Chem. Soc **128**(31): 10004-10005.

Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus (1983). "CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations." Journal of Computational Chemistry **4**(2): 187-217.

Bucciantini, M., E. Giannoni, F. Chiti, F. Baroni, L. Formigli, J. Zurdo, N. Taddei, G. Ramponi, C. M. Dobson and M. Stefani (2002). "Inherent Toxicity of Aggregates Implies a Common Mechanism for Protein Misfolding Diseases." Nature **416**(6880): 507-511.

Bussell, R. and D. Eliezer (2003). "A Structural and Functional Role for 11-mer Repeats in α-Synuclein and Other Exchangeable Lipid Binding Proteins." Journal of molecular biology **329**(4): 763-778.

Bystroff, C. and S. Garde (2003). "Helix Propensities of Short Peptides: Molecular Dynamics Versus Bioinformatics." Proteins **50**(4): 552-562.

Chandra, S., X. Chen, J. Rizo, R. Jahn and T. C. Sudhof (2003). "Protein Structure and Folding - A Broken a-Helix in Folded α-Synuclein." Journal of Biological Chemistry **278**(17): 15313-15318.

Chen, L. Y. and N. J. Horing (2007). "An Exact Formulation of Hyperdynamics Simulations." J Chem Phys **126**(22): 224103.

Chen, Y., S. L. Campbell and N. V. Dokholyan (2007). "Deciphering Protein Dynamics from NMR Data Using Explicit Structure Sampling and Selection." Biophysical journal **93**(7): 2300-2306.

Cho, M.-K., G. Nodet, H.-Y. Kim, M. R. Jensen, P. Bernado, C. O. Fernandez, S. Becker, M. Blackledge and M. Zweckstetter (2009). "Structural Characterization of α-Synuclein in an Aggregation Prone State." Protein Science **18**(9): 1840-1846.

Choy, W. Y. and J. D. Forman-Kay (2001). "Calculation of Ensembles of Structures Representing the Unfolded State of an SH3 Domain." Journal of molecular biology **308**(5): 1011-1032.

Conway, K. A., S.-J. Lee, J.-C. Rochet, T. T. Ding, R. E. Williamson and P. T. Lansbury (2000). "Acceleration of Oligomerization, not Fibrillization, is a Shared Property of both α-Synuclein Mutations Linked to Early-Onset Parkinson's Disease: Implications for Pathogenesis and Therapy." Proceedings of the National Academy of Sciences of the United States of America **97**(2): 571-576.

Cooper, A. A., A. D. Gitler, A. Cashikar, C. M. Haynes, K. J. Hill, B. Bhullar, K. Liu, K. Xu, K. E. Strathearn, F. Liu, S. Cao, K. A. Caldwell, G. A. Caldwell, G. Marsischky, R. D. Kolodner, J. LaBaer, J.-C. Rochet, N. M. Bonini and S. Lindquist (2006). "α-Synuclein Blocks ER-Golgi Traffic and Rab1 Rescues Neuron Loss in Parkinson's Models." Science **313**(5785): 324-328.

Crabtree, D. M. and J. Zhang (2012). "Genetically Engineered Mouse Models of Parkinson's Disease." Brain Res Bull **88**(1): 13-32.

Damaschun, G., H. Damaschun, K. Gast, R. Misselwitz, J. J. Muller, W. Pfeil and D. Zirwer (1993). "Cold Denaturation-Induced Conformational Changes in Phosphoglycerate Kinase from Yeast." Biochemistry **32**(30): 7739-7746.

Danzer, K. M., D. Haasen, A. R. Karow, S. Moussaud, M. Habeck, A. Giese, H. Kretzschmar, B. Hengerer and M. Kostka (2007). "Different Species of α-Synuclein Oligomers Induce Calcium Influx and Seeding." The Journal of Neuroscience **27**(34): 9220-9232.

Davidson, W. S., A. Jonas, D. F. Clayton and J. M. George (1998). "Stabilization of α-Synuclein Secondary Structure upon Binding to Synthetic Membranes." Journal of Biological Chemistry **273**(16): 9443-9449.

Dedmon, M. M., K. Lindorff-Larsen, J. Christodoulou, M. Vendruscolo and C. M. Dobson (2005). "Mapping Long-Range Interactions in α-Synuclein Using Spin-Label NMR and Ensemble Molecular Dynamics Simulations." Journal of the American Chemical Society **127**(2): 476-477.

Drescher, M., M. Huber and V. Subramaniam (2012). "Hunting the Chameleon: Structural Conformations of the Intrinsically Disordered Protein α-Synuclein." ChemBioChem **13**(6): 761-768.

Dudzik, C. G., E. D. Walter and G. L. Millhauser (2011). "Coordination Features and Affinity of the Cu^{2+} Site in the α-Synuclein Protein of Parkinson's Disease." Biochemistry **50**(11): 1771-1777.

Dzubiella, J. (2009). "Sequence-Specific Size, Structure, and Stability of Tight Protein Knots." Biophys J **96**(3): 831-839.

El-Agnaf, O. M. and G. B. Irvine (2002). "Aggregation and Neurotoxicity of α-Synuclein and Related Peptides." Biochemical Society transactions **30**(4): 559-565.

Eliezer, D., J. Chung, H. J. Dyson and P. E. Wright (2000). "Native and Non-Native Secondary Structure and Dynamics in the pH 4 Intermediate of Apomyoglobin." Biochemistry **39**(11): 2894-2901.

Eliezer, D., E. Kutluay, R. Bussell and G. Browne (2001). "Conformational Properties of α-Synuclein in its Free and Lipid-Associated States." Journal of molecular biology **307**(4): 1061-1073.

Esteban-Martin, S., R. B. Fenwick and X. Salvatella (2009). "Refinement of Ensembles Describing Unstructured Proteins using NMR Residual Dipolar Couplings." J Am Chem Soc **132**(13): 4626-4632.

Fauvet, B., M. M. Kamdem, M.B. Fares, C. Desobry, S. Michael, M. T. Ardah, E. Tsika, P. Coune, M. Prudent, N. Lion, D. Eliezer, D. J. Moore, B. Schneider, P. Aebischer, O. M. El-Agnaf, E. Masliah and H. A. Lashuel (2012). "α-Synuclein in the Central Nervous System and from Erythrocytes, Mammalian Cells and *E. coli* Exists Predominantly as a Disordered Monomer." Journal of Biological Chemistry.

Fawzi, N. L., A. H. Phillips, J. Z. Ruscio, M. Doucleff, D. E. Wemmer and T. Head-Gordon (2008). "Structure and Dynamics of the $A\beta_{21-30}$ Peptide from the Interplay of NMR Experiments and Molecular Simulations." Journal of the American Chemical Society **130**(19): 6145-6158.

Feldman, H. J. and C. W. V. Hogue (2000). "A Fast Method to Sample Real Protein Conformational Space." Proteins: Structure, Function, and Bioinformatics **39**(2): 112-131.

Ferreon, A. C. M., Y. Gambin, E. A. Lemke and A. A. Deniz (2009). "Interplay of α-Synuclein Binding and Conformational Switching Probed by Single-Molecule Fluorescence." Proceedings of the National Academy of Sciences **106**(14): 5645.

Fink, A. L. (2006). "The Aggregation and Fibrillation of α-Synuclein." Accounts of Chemical Research **39**(9): 628-634.

Fisher, C. K., A. Huang and C. M. Stultz (2010). "Modeling Intrinsically Disordered Proteins with Bayesian Statistics." Journal of the American Chemical Society **132**(42): 14919-14927.

Fisher, C. K. and C. M. Stultz (2011). "Constructing Ensembles for Intrinsically Disordered Proteins." Current Opinion in Structural Biology **21**(3): 426-431.

Fisher, C. K. and C. M. Stultz (2011). "Protein Structure along the Order-Disorder Continuum." J Am Chem Soc **133**(26): 10022-10025.

Fisher, C. K., Ullman, O., and Stultz, C.M. (2012). "Efficient Construction of Disordered Protein Ensembles in a Bayesian Framework with Optimal Selection of Conformations." Pacific Symposium on Biocomputing **17**: 82-93.

Flory, P. J. (1953). Principles of Polymer Chemistry, Cornell University Press.

Flory, P. J. and S. Fisk (1966). "Effect of Volume Exclusion on the Dimensions of Polymer Chains." The Journal of Chemical Physics **44**(6): 2243-2248.

Forno, L. S. (1996). "Neuropathology of Parkinson's Disease." Journal of Neuropathology and Experimental Neurology **55**(3): 259-272.

Frimpong, A. K., R. R. Abzalimov, V. N. Uversky and I. A. Kaltashov (2010). "Characterization of Intrinsically Disordered Proteins with Electrospray Ionization Mass Spectrometry: Conformational Heterogeneity of α-Synuclein." Proteins: Structure, Function, and Bioinformatics **78**(3): 714-722.

Frishman, D. and P. Argos (1995). "Knowledge-Based Protein Secondary Structure Assignment." Proteins: Structure, Function, and Bioinformatics **23**(4): 566-579.

Galvin, J. E., V. M. Y. Lee and J. Q. Trojanowski (2001). "Synucleinopathies: Clinical and Pathological Implications." Archives of neurology **58**(2): 186.

Gast, K., H. Damaschun, K. Eckert, K. Schulze-Forster, H. R. Maurer, M. Mueller-Frohne, D. Zirwer, J. Czarnecki and G. Damaschun (1995). "Prothymosin a: A Biologically Active Protein with Random Coil Conformation." Biochemistry **34**(40): 13211-13218.

Gast, K., H. Damaschun, K. Eckert, K. Schulze-Forster, H. R. Maurer, M. Muller-Frohne, D. Zirwer, J. Czarnecki and G. Damaschun (1995). "Prothymosin a: A Biologically Active Protein with Random Coil Conformation." Biochemistry **34**(40): 13211-13218.

George, J. M., H. Jin, W. S. Woods and D. F. Clayton (1995). "Characterization of a Novel Protein Regulated During the Critical Period for Song Learning in the Zebra Finch." Neuron **15**(2): 361-372.

Georgieva, E. R., T. F. Ramlall, P. P. Borbat, J. H. Freed and D. Eliezer (2008). "Membrane-Bound α-Synuclein Forms an Extended Helix: Long-Distance Pulsed ESR Measurements Using Vesicles, Bicelles, and Rodlike Micelles." Journal of the American Chemical Society **130**(39): 12856-12857.

Georgieva, E. R., T. F. Ramlall, P. P. Borbat, J. H. Freed and D. Eliezer (2010). "The Lipid-binding Domain of Wild Type and Mutant α-Synuclein." Journal of Biological Chemistry **285**(36): 28261.

Gillespie, J. R. and D. Shortle (1997). "Characterization of Long-Range Structure in the Denatured State of Staphylococcal Nuclease .I. Paramagnetic Relaxation Enhancement by Nitroxide Spin Labels." Journal of molecular biology **268**(1): 158-169.

Goddard, T. D. and D. G. Kneller SPARKY 3, University of California, San Francisco.

Goedert, M. (2001). "α-Synuclein and Neurodegenerative Diseases." Nature Reviews Neuroscience **2**(7): 492-501.

Greenfield, N. J. (2007). "Using Circular Dichroism Spectra to Estimate Protein Secondary Structure." Nat. Protocols **1**(6): 2876-2890.

Heinig, M. and D. Frishman (2004). "STRIDE: a Web Server for Secondary Structure Assignment from Known Atomic Coordinates of Proteins." Nucleic Acids Research **32**: W500-W502.

Ho, B. K. and K. A. Dill (2006). "Folding Very Short Peptides Using Molecular Dynamics." Plos Computational Biology **2**(4): 228-237.

Huang, A. and C. M. Stultz (2007). "Conformational Sampling with Implicit Solvent Models: Application to the PHF6 Peptide in Tau Protein." Biophys J **92**(1): 34-45.

Huang, A. and C. M. Stultz (2008). "The Effect of a ΔK280 Mutation on the Unfolded State of a Microtubule-Binding Repeat in Tau." Plos Computational Biology **4**(8): 12.

Huang, A. and C. M. Stultz (2009). "Finding Order within Disorder: Elucidating the Structure of Proteins Associated with Neurodegenerative Disease." Future Medicinal Chemistry **1**(3): 467-482.

Iwai, A., E. Masliah, M. Yoshimoto, N. Ge, L. Flanagan, H. A. Rohan de Silva, A. Kittel and T. Saitoh (1995). "The Precursor Protein of Non-Aβ Component of Alzheimer's Disease Amyloid Is a Presynaptic Protein of the Central Nervous System." Neuron **14**(2): 467-475.

Iwai, A., M. Yoshimoto, E. Masliah and T. Saitoh (1995). "Non-Aβ Component of Alzheimer's Disease Amyloid (NAC) is Amyloidogenic." Biochemistry **34**(32): 10139-10145.

Jao, C. C., B. G. Hegde, J. Chen, I. S. Haworth and R. Langen (2008). "Structure of Membrane-Bound α-Synuclein from Site-Directed Spin Labeling and Computational Refinement." Proceedings of the National Academy of Sciences of the United States of America **105**(50): 19666.

Jarrett, J. T. and P. T. Lansbury (1993). "Seeding One-Dimensional Crystallization of Amyloid - a Pathogenic Mechanism in Alzheimers-Disease and Scrapie." Cell **73**(6): 1055-1058.

Jensen, M. R., L. Salmon, G. Nodet and M. Blackledge (2010). "Defining Conformational Ensembles of Intrinsically Disordered and Partially Folded Proteins Directly from Chemical Shifts." J Am Chem Soc **132**(4): 1270-1272.

Jha, A. K., A. Colubri, K. F. Freed and T. R. Sosnick (2005). "Statistical Coil Model of the Unfolded State: Resolving the Reconciliation Problem." Proceedings of the National Academy of Sciences of the United States of America **102**(37): 13099-13104.

Kabsch, W. and C. Sander (1983). "Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features." Biopolymers **22**(12): 2577-2637.

Kamtekar, S. and M. H. Hecht (1995). "Protein Motifs .7. The 4-Helix Bundle - What Determines a Fold." Faseb Journal **9**(11): 1013-1022.

Kang, L., G. M. Moriarty, L. A. Woods, A. E. Ashcroft, S. E. Radford and J. Baum (2012). "N-Terminal Acetylation of α-Synuclein Induces Increased Transient Helical Propensity and Decreased Aggregation Rates in the Intrinsically Disordered Monomer." Protein Sci **21**(7): 911-917.

Karplus, M. and J. A. McCammon (2002). "Molecular Dynamics Simulations of Biomolecules." Nat Struct Biol **9**(9): 646-652.

Kay, L., P. Keifer and T. Saarinen (1992). "Pure Absorption Gradient Enhanced Heteronuclear Single Quantum Correlation Spectroscopy with Improved Sensitivity." Journal of the American Chemical Society **114**(26): 10663-10665.

Kayed, R., E. Head, J. L. Thompson, T. M. McIntire, S. C. Milton, C. W. Cotman and C. G. Glabe (2003). "Common Structure of Soluble Amyloid Oligomers Implies Common Mechanism of Pathogenesis." Science **300**(5618): 486-489.

King, N. P., A. W. Jacobitz, M. R. Sawaya, L. Goldschmidt and T. O. Yeates (2010). "Structure and Folding of a Designed Knotted Protein." Proceedings of the National Academy of Sciences of the United States of America **107**(48): 20732-20737.

Kohn, J., I. Millett, J. Jacob, B. Zagrovic, T. Dillon, N. Cingel, R. Dothager, S. Seifert, P. Thiyagarajan and T. Sosnick (2005). "Correction: Random-Coil Behavior and the Dimensions of Chemically Unfolded Proteins (vol 101, pg 12491, 2004)." Proceedings of the National Academy of Sciences of the United States of America **102**(40): 14475-14475.

Kohn, J. E., I. S. Millett, J. Jacob, B. Zagrovic, T. M. Dillon, N. Cingel, R. S. Dothager, S. Seifert, P. Thiyagarajan, T. R. Sosnick, M. Z. Hasan, V. S. Pande, I. Ruczinski, S. Doniach and K. W. Plaxco (2004). "Random-Coil Behavior and the Dimensions of Chemically Unfolded Proteins." Proceedings of the National Academy of Sciences of the United States of America **101**(34): 12491-12496.

Kolesov, G., P. Virnau, M. Kardar and L. A. Mirny (2007). "Protein Knot Server: Detection of Knots in Protein Structures." Nucleic Acids Research **35**: W425-W428.

Koo, H. J., M. Y. Choi and H. Im (2009). "Aggregation-Defective α-Synuclein Mutants Inhibit the Fibrillation of Parkinson's Disease-Linked α-Synuclein Variants." Biochemical and Biophysical Research Communications **386**(1): 165-169.

Laganowsky, A., C. Liu, M. R. Sawaya, J. P. Whitelegge, J. Park, M. L. Zhao, A. Pensalfini, A. B. Soriaga, M. Landau, P. K. Teng, D. Cascio, C. Glabe and D. Eisenberg (2012). "Atomic View of a Toxic Amyloid Small Oligomer." Science **335**(6073): 1228-1231.

Lane, T. J., C. R. Schwantes, K. A. Beauchamp and V. S. Pande (2014). "Efficient Inference of Protein Structural Ensembles." arXiv preprint arXiv:1408.0255.

Lawson, D. M., P. J. Artymiuk, S. J. Yewdall, J. M. A. Smith, J. C. Livingstone, A. Treffry, A. Luzzago, S. Levi, P. Arosio, G. Cesareni, C. D. Thomas, W. V. Shaw and P. M. Harrison (1991). "Solving the Structure of Human H Ferritin by Genetically Engineering Intermolecular Crystal Contacts." Nature **349**(6309): 541-544.

Lazaridis, T. and M. Karplus (1999). "Effective Energy Function for Proteins in Solution." Proteins-Structure Function and Genetics **35**(2): 133-152.

Lesage, S., M. Anheim, F. Letournel, L. Bousset, A. Honore, N. Rozas, L. Pieri, K. Madiona, A. Durr, R. Melki, C. Verny, A. Brice and G. French Parkinson's Disease Genetics Study (2013). "G51D α-Synuclein Mutation Causes a Novel Parkinsonian-Pyramidal Syndrome." Ann Neurol **73**(4): 459-471.

Lewy, F. (1912). "Paralysis Agitans. I. Pathologische Anatomie." Handbuch der neurologie **3**(part 2): 920-933.

Li, J., V. N. Uversky and A. L. Fink (2001). "Effect of Familial Parkinson's Disease Point Mutations A30P and A53T on the Structural Properties, Aggregation, and Fibrillation of Human α-Synuclein." Biochemistry **40**(38): 11604-11613.

Liu, J., N. B. Perumal, C. J. Oldfield, E. W. Su, V. N. Uversky and A. K. Dunker (2006). "Intrinsic Disorder in Transcription Factors " Biochemistry **45**(22): 6873-6888.

Low, K. and P. Aebischer (2012). "Use of Viral Vectors to Create Animal Models for Parkinson's Disease." Neurobiol Dis **48**(2): 189-201.

Manavalan, P. and W. C. Johnson (1985). "Protein Secondary Structure from Circular-Dichroism Spectra." Journal of Biosciences **8**(1-2): 141-149.

Maries, E., B. Dass, T. J. Collier, J. H. Kordower and K. Steece-Collier (2003). "The Role of α-Synuclein in Parkinson's Disease: Insights from Animal Models." Nat Rev Neurosci **4**(9): 727-738.

Marqusee, S., V. H. Robbins and R. L. Baldwin (1989). "Unusually Stable Helix Formation in Short Alanine-Based Peptides." Proc Natl Acad Sci U S A **86**(14): 5286-5290.

Marsh, J. A., J. M. R. Baker, M. Tollinger and J. D. Forman-Kay (2008). "Calculation of Residual Dipolar Couplings from Disordered State Ensembles Using Local Alignment." Journal of the American Chemical Society **130**(25): 7804-+.

Marsh, J. A. and J. D. Forman-Kay (2009). "Structure and Disorder in an Unfolded State under Nondenaturing Conditions from Ensemble Models Consistent with a Large Number of Experimental Restraints." Journal of molecular biology **391**(2): 359-374.

Mathews, F. S., P. H. Bethge and E. W. Czerwinski (1979). "The Structure of Cytochrome b_{562} from *Escherichia coli* at 2.5 Å Resolution." J Biol Chem **254**(5): 1699-1706.

Merelo, J. J., M. A. Andrade, A. Prieto and F. Moran (1994). "Proteinotopic Feature Maps." Neurocomputing **6**(4): 443-454.

Mittag, T. and J. D. Forman-Kay (2007). "Atomic-Level Characterization of Disordered Protein Ensembles." Current Opinion in Structural Biology **17**(1): 3-14.

Mukherjea, M., P. Llinas, H. Kim, M. Travaglia, D. Safer, J. Ménétrey, C. Franzini-Armstrong, P. R. Selvin, A. Houdusse and H. L. Sweeney (2009). "Myosin VI Dimerization Triggers an Unfolding of a Three-Helix Bundle in Order to Extend Its Reach." Molecular cell **35**(3): 305-315.

Neal, S., A. M. Nip, H. Zhang and D. S. Wishart (2003). "Rapid and Accurate Calculation of Protein ^{1}H, ^{13}C and ^{15}N Chemical Shifts." Journal of Biomolecular NMR **26**(3): 215-240.

Neria, E., S. Fischer and M. Karplus (1996). "Simulation of Activation Free Energies in Molecular Systems." Journal of Chemical Physics **105**(5): 1902-1921.

Nodet, G., L. Salmon, V. Ozenne, S. Meier, M. R. Jensen and M. Blackledge (2009). "Quantitative Description of Backbone Conformational Sampling of Unfolded Proteins at Amino Acid Resolution from NMR Residual Dipolar Couplings." Journal of the American Chemical Society **131**(49): 17908-17918.

Ozenne, V., F. Bauer, L. Salmon, J. R. Huang, M. R. Jensen, S. Segard, P. Bernado, C. Charavay and M. Blackledge (2012). "Flexible-Meccano: A Tool for the Generation of Explicit Ensemble Descriptions of Intrinsically Disordered Proteins and their Associated Experimental Observables." Bioinformatics **28**(11): 1463-1470.

Pitera, J. W. and J. D. Chodera (2012). "On the Use of Experimental Observations to Bias Simulated Ensembles." Journal of Chemical Theory and Computation **8**(10): 3445-3451.

Radivojac, P., L. M. Iakoucheva, C. J. Oldfield, Z. Obradovic, V. N. Uversky and A. K. Dunker (2007). "Intrinsic Disorder and Functional Proteomics." Biophysical Journal **92**(5): 1439-1456.

Rao, J. N., Y. E. Kim, L. S. Park and T. S. Ulmer (2009). "Effect of Pseudorepeat Rearrangement on α-Synuclein Misfolding, Vesicle Binding, and Micelle Binding." Journal of molecular biology **390**(3): 516-529.

Rauscher, S. and R. Pomes (2010). "Molecular Simulations of Protein Disorder." Biochem Cell Biol **88**(2): 269-290.

Rekas, A., R. B. Knott, A. Sokolova, K. J. Barnham, K. A. Perez, C. L. Masters, S. C. Drew, R. Cappai, C. C. Curtain and C. L. L. Pham (2010). "The Structure of Dopamine Induced α-Synuclein Oligomers." European Biophysics Journal with Biophysics Letters **39**(10): 1407-1419.

Rospigliosi, C. C., S. McClendon, A. W. Schmid, T. F. Ramlall, P. Barre, H. A. Lashuel and D. Eliezer (2009). "E46K Parkinson's-Linked Mutation Enhances C-Terminal-to-N-Terminal Contacts in α-Synuclein." Journal of molecular biology **388**(5): 1022-1032.

Sandal, M., F. Valle, I. Tessari, S. Mammi, E. Bergantino, F. Musiani, M. Brucale, L. Bubacco and B. Samorì (2008). "Conformational Equilibria in Monomeric α-Synuclein at the Single-Molecule Level." PLoS Biol **6**(1): 0099-0108.

Schrodinger, LLC (2010). The PyMOL Molecular Graphics System, Version 1.3r1.

Schwalbe, H., K. M. Fiebig, M. Buck, J. A. Jones, S. B. Grimshaw, A. Spencer, S. J. Glaser, L. J. Smith and C. M. Dobson (1997). "Structural and Dynamical Properties of a Denatured Protein.

Heteronuclear 3D NMR Experiments and Theoretical Simulations of Lysozyme in 8M Urea." Biochemistry **36**(29): 8977-8991.

Schwieters, C. D., J. J. Kuszewski, N. Tjandra and G. Marius Clore (2003). "The Xplor-NIH NMR Molecular Structure Determination Package." Journal of Magnetic Resonance **160**(1): 65-73.

Serpell, L. C., J. Berriman, R. Jakes, M. Goedert and R. A. Crowther (2000). "Fiber Diffraction of Synthetic α-Synuclein Filaments Shows Amyloid-Like Cross-β Conformation." Proceedings of the National Academy of Sciences of the United States of America **97**(9): 4897.

Sgourakis, N. G., M. Merced-Serrano, C. Boutsidis, P. Drineas, Z. M. Du, C. Y. Wang and A. E. Garcia (2011). "Atomic-Level Characterization of the Ensemble of the Aβ(1-42) Monomer in Water Using Unbiased Molecular Dynamics Simulations and Spectral Algorithms." Journal of Molecular Biology **405**(2): 570-583.

Shin, H. C., G. Merutka, J. P. Waltho, L. L. Tennant, H. J. Dyson and P. E. Wright (1993). "Peptide Models of Protein Folding Initiation Sites. 3. The G-H helical Hairpin of Myoglobin." Biochemistry **32**(25): 6356-6364.

Sickmeier, M., J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic and A. K. Dunker (2007). "DisProt: the Database of Disordered Proteins." Nucleic Acids Res **35**(Database issue): D786-793.

Spillantini, M. G., R. A. Crowther, R. Jakes, M. Hasegawa and M. Goedert (1998). "α-Synuclein in Filamentous Inclusions of Lewy Bodies from Parkinson's Disease and Dementia with Lewy Bodies." Proceedings of the National Academy of Sciences of the United States of America **95**(11): 6469-6473.

Spillantini, M. G. and M. Goedert (2000). "The α-Synucleinopathies: Parkinson's Disease, Dementia with Lewy Bodies, and Multiple System Atrophy." Annals of the New York Academy of Sciences **920**(THE MOLECULAR BASIS OF DEMENTIA): 16-27.

Spillantini, M. G., M. L. Schmidt, V. M. Y. Lee, J. Q. Trojanowski, R. Jakes and M. Goedert (1997). "α-Synuclein in Lewy Bodies." Nature **388**(6645): 839-840.

Steinbach, P. J. (2004). "Exploring Peptide Energy Landscapes: A Test of Force Fields and Implicit Solvent Models." Proteins **57**(4): 665-677.

Strodel, B. and D. J. Wales (2008). "Implicit Solvent Models and the Energy Landscape for Aggregation of the Amyloidogenic KFFE Peptide." Journal of Chemical Theory and Computation **4**(4): 657-672.

Stultz, C. M., J. V. White and T. F. Smith (1993). "Structural-Analysis Based on State-Space Modeling." Protein Science **2**(3): 305-314.

Sugita, Y. and Y. Okamoto (1999). "Replica-Exchange Molecular Dynamics Method for Protein Folding." Chemical Physics Letters **314**(1-2): 141-151.

Sung, Y. and D. Eliezer (2007). "Residual Structure, Backbone Dynamics, and Interactions within the Synuclein Family." Journal of molecular biology **372**(3): 689-707.

Trexler, A. J. and E. Rhoades (2009). "α-Synuclein Binds Large Unilamellar Vesicles as an Extended Helix." Biochemistry **48**(11): 2304-2306.

Trexler, A. J. and E. Rhoades (2012). "N-Terminal Acetylation is Critical for Forming α-Helical Oligomer of α-Synuclein." Protein Science **21**(5): 601-605.

Ullman, O., C. K. Fisher and C. M. Stultz (2011). "Explaining the Structural Plasticity of α-Synuclein." Journal of the American Chemical Society **133**(48): 19536-19546.

Ulmer, T. S., A. Bax, N. B. Cole and R. L. Nussbaum (2005). "Structure and Dynamics of Micelle-Bound Human α-Synuclein." Journal of Biological Chemistry **280**(10): 9595-9603.

Uversky, V. N. (2002). "Natively Unfolded Proteins: A Point Where Biology Waits for Physics." Protein Science **11**(4): 739-756.

Uversky, V. N. (2003). "A Protein-Chameleon: Conformational Plasticity of α-Synuclein, a Disordered Protein Involved in Neurodegenerative Disorders." Journal of Biomolecular Structure and Dynamics **21**(2): 159-310.

Uversky, V. N., J. R. Gillespie and A. L. Fink (2000). "Why are "Natively Unfolded" Proteins Unstructured under Physiologic Conditions?" Proteins: Structure, Function, and Bioinformatics **41**(3): 415-427.

Uversky, V. N., J. Li and A. L. Fink (2001). "Evidence for a Partially Folded Intermediate in α-Synuclein Fibril Formation." Journal of Biological Chemistry **276**(14): 10737-10744.

Uversky, V. N., C. J. Oldfield and A. K. Dunker (2008). "Intrinsically Disordered Proteins in Human Diseases: Introducing the D^2 Concept." Annual Review of Biophysics **37**: 215-246.

Vamvaca, K., M. J. Volles and P. T. Lansbury (2009). "The First N-terminal Amino Acids of α-Synuclein Are Essential for a-Helical Structure Formation In Vitro and Membrane Binding in Yeast." Journal of Molecular Biology **389**(2): 413-424.

Venda, L. L., S. J. Cragg, V. L. Buchman and R. Wade-Martins (2010). "α-Synuclein and Dopamine at the Crossroads of Parkinson's Disease." Trends Neurosci **33**(12): 559-568.

Virnau, P., L. A. Mirny and M. Kardar (2006). "Intricate Knots in Proteins: Function and Evolution." Plos Computational Biology **2**(9): 1074-1079.

Voelz, V. A., M. S. Shell and K. A. Dill (2009). "Predicting Peptide Structures in Native Proteins from Physical Simulations of Fragments." PLoS Comput Biol **5**(2): e1000281.

Volles, M. J., S. J. Lee, J. C. Rochet, M. D. Shtilerman, T. T. Ding, J. C. Kessler and P. T. Lansbury, Jr. (2001). "Vesicle Permeabilization by Protofibrillar α-Synuclein: Implications for the Pathogenesis and Treatment of Parkinson's Disease." Biochemistry **40**(26): 7812-7819.

von Hippel, P. H. (2004). "Completing the View of Transcriptional Regulation." Science **305**(5682): 350-352.

Waltho, J. P., V. A. Feher, G. Merutka, H. J. Dyson and P. E. Wright (1993). "Peptide Models of Protein Folding Initiation Sites. 1. Secondary Structure Formation by Peptides Corresponding to the G- and H-Helices of Myoglobin." Biochemistry **32**(25): 6337-6347.

Wang, W., I. Perovic, J. Chittuluru, A. Kaganovich, L. T. T. Nguyen, J. Liao, J. R. Auclair, D. Johnson, A. Landeru, A. K. Simorellis, S. Ju, M. R. Cookson, F. J. Asturias, J. N. Agar, B. N. Webb, C. Kang, D. Ringe, G. A. Petsko, T. C. Pochapsky and Q. Q. Hoang (2011). "A Soluble α-Synuclein Construct Forms a Dynamic Tetramer." Proceedings of the National Academy of Sciences of the United States of America **108**(43): 17797-17802.

Weinreb, P. H., W. G. Zhen, A. W. Poon, K. A. Conway and P. T. Lansbury (1996). "NACP, a Protein Implicated in Alzheimer's Disease and Learning, is Natively Unfolded." Biochemistry **35**(43): 13709-13715.

Whitworth, A. J. (2011). "Drosophila Models of Parkinson's Disease." Adv Genet **73**: 1-50.

Winner, B., R. Jappelli, S. K. Maji, P. A. Desplats, L. Boyer, S. Aigner, C. Hetzer, T. Loher, M. a. Vilar, S. Campioni, C. Tzitzilonis, A. Soragni, S. Jessberger, H. Mira, A. Consiglio, E. Pham, E. Masliah, F. H. Gage and R. Riek (2011). "In Vivo Demonstration that α-Synuclein Oligomers are Toxic." Proceedings of the National Academy of Sciences of the United States of America **108**(10): 4194-4199.

Wishart, D., C. Bigam, A. Holm, R. Hodges and B. Sykes (1995). "^{1}H, ^{13}C and ^{15}N Random Coil NMR Chemical Shifts of the Common Amino Acids. I. Investigations of Nearest-Neighbor Effects." Journal of Biomolecular NMR **5**(1): 67-81.

Wu, K.-P., S. Kim, D. A. Fela and J. Baum (2008). "Characterization of Conformational and Dynamic Properties of Natively Unfolded Human and Mouse α-Synuclein Ensembles by NMR: Implication for Aggregation." Journal of Molecular Biology **378**(5): 1104-1115.

Wu, K.-P., D. S. Weinstock, C. Narayanan, R. M. Levy and J. Baum (2009). "Structural Reorganization of α-Synuclein at Low pH Observed by NMR and REMD Simulations." J Mol Biol **391**(4): 784-796.

Zerella, R., P. A. Evans, J. M. Ionides, L. C. Packman, B. W. Trotter, J. P. Mackay and D. H. Williams (1999). "Autonomous Folding of a Peptide Corresponding to the N-terminal β-hairpin from Ubiquitin." Protein Sci **8**(6): 1320-1331.

Zhou, H.-X. (2001). "The Affinity-Enhancing Roles of Flexible Linkers in Two-Domain DNA-Binding Proteins " Biochemistry **40**(50): 15069-15073.
Zweckstetter, M. (2008). "NMR: Prediction of Molecular Alignment from Structure Using the PALES Software." Nat. Protocols **3**(4): 679-690.

Zweckstetter, M. and A. Bax (2000). "Prediction of Sterically Induced Alignment in a Dilute Liquid Crystalline Phase: Aid to Protein Structure Determination by NMR." J. Am. Chem. Soc **122**(15): 3791-3792.

Orly Ullman

Massachusetts Institute of Technology
77 Massachusetts Ave, E25-530
Cambridge, MA 02139
617.999.6792 | orly@mit.edu

Education

2008-2015 Ph.D. in Physical Chemistry, Massachusetts Institute of Technology (MIT)
Thesis title: "Building Models of Intrinsically Disordered Protein;
a Comprehensive Study of α-Synuclein", Advisor: Prof. Collin M. Stultz
2005-2007 Graduate studies in the Chemistry Department, Hebrew University, Israel
Supervisor, Prof. Avinoam Ben-Shaul, Dept. of Physical Chemistry
Used mean-field theories and Monte-Carlo simulations to study the formation of junctions in the adhesion areas between adjacent cells
2002-2005 Undergraduate studies: Tel-Aviv University, Israel
B.Sc. in Chemistry, Graduated with honors
1994-2000 De Shalit High School, Israel
Graduated with honors, Full Matriculation diploma,
Enhanced studies in Physics, Chemistry, English and Math

Professional and Teaching Experience

Fall 2008, Teaching Assistant for 5.111 (Principles of Chemical Science), MIT
Conducted biweekly recitation sections, graded homework and exams, and maintained grade records.
Spring 2009, Teaching Assistant for 5.60 (Thermodynamics and Kinetics), MIT
Same responsibilities as above.
Summer 2005, "Advanced Chemistry Workshop", Tel Aviv University
Conducted a research project under the supervision of Prof. Israel Goldberg. Project entailed the study of crystals using X-ray diffraction, finding their structure and improving results.
Summer 2004, Emma and Oscar Gets Summer Science Program for honorary students at the Weizmann Institute, Israel
Took part in research titled "The Formation of Alkylsilanes Monolayer Using Two Preparation Techniques" under the supervision of Prof. Ron Naaman.

Awards & Honors

2012
Elie Shaio Memorial Award, MIT
2005-2006
Klein's Scholarship of Excellence for graduate students, the Hebrew University, Israel

2004

Emma and Oscar Gets Summer Science Program for honorary students at the Weizmann Institute, Israel

Posters and Presentations

2012, Pacific Symposium on Biocomputing

Poster titled: "Efficient Construction of Disordered Protein Ensembles in a Bayesian Framework with Optimal Selection of Conformations"

2011, Biophysical Society, 55th annual meeting

Presented a poster titled: "Modeling α-Synuclein Structural Ensemble Using a Bayesians Weighting Approach"

The poster was identified as being a particularly important development in the field by a member of F1000 and was invited for evaluation

Summer 2007, Physics of biological matter, research workshop, Israel

Presented a poster titled: "Do Thermodynamic Principles Play a Role in Intracellular Connections and Tissue Organization"

Publications

Thomas Gurry*, Orly Ullman* ,Charles K Fisher, Iva Perovic, Thomas Pochapsky, Collin M. Stultz "**The Dynamic Structure of α-Synuclein Multimers**" *J. Am Chem Soc.* 2013, 135(10): 3865-3872

Charles K Fisher, Orly Ullman, Collin M Stultz "**Comparative Studies of Disordered Proteins with Similar Sequences: Application to Aβ40 and Aβ42**" *Biophysical J.* 2013, 104(7): 1546 – 1555

Sophie Walker, Orly Ullman, Collin M Stultz "**Using Intramolecular Disulfide Bonds in Tau Protein to Deduce Structural Features of Aggregation-resistant Conformations**" *J. Bio. Chem.* 2012, 287(12): 9591-9600

Charles K Fisher, Orly Ullman, Collin M Stultz "**Efficient Construction of Disordered Protein Ensembles in a Bayesian Framework with Optimal Selection of Conformations**" *Pac. Symp. Biocomput.* 2012, 82-93

Orly Ullman, Charles K Fisher, Collin M Stultz "**Explaining the Structural Plasticity of α-Synuclein**" *J. Am Chem Soc.* 2011, 133(48): 19536-19546

*Both authors contributed equally to this work

Other Service

2000-2002

Full mandatory service and three months standing army service as Non-Commissioned Officer in elite Military Intelligence unit 8200. Role included data acquisition and management in specific field of expertise, serving as professional authority and giving lectures to various elements in the unit and outside it. Awarded for outstanding performance.