

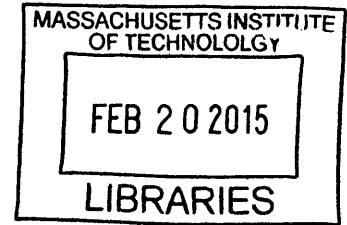
Order, Disorder, and Protein Aggregation

By

Thomas Gurry

B.Sc. Imperial College London, 2008
M.Phil. University of Cambridge, 2009

ARCHIVES



SUBMITTED TO THE COMPUTATIONAL AND SYSTEMS BIOLOGY PROGRAM IN
FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN COMPUTATIONAL AND SYSTEMS BIOLOGY
AT THE MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2015

© 2015 Massachusetts Institute of Technology. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly
paper and electronic copies of this thesis document in whole or in part in any medium
now known or hereafter created.

Signature redacted

Signature of the Author: _____
Computational and Systems Biology Ph.D. Program
January 14, 2015

Signature redacted

Certified by: _____
Collin M. Stultz
Professor of Electrical Engineering and Computer Science
and Health Sciences and Technology
Thesis Supervisor
Signature redacted

Accepted by: _____
Christopher B. Burge
Professor of Biology and Biological Engineering
Director, Computational and Systems Biology Ph.D. Program

Order, Disorder, and Protein Aggregation

by

Thomas Gurry

Submitted to the Computational and Systems Biology Program
on January 15, 2015 in Fulfillment of
the Requirements for the degree of
Doctor of Philosophy in Computational and Systems Biology

ABSTRACT

Protein aggregation underlies a number of human diseases. Most notably, it occurs widely in neurodegenerative diseases, including Alzheimer's and Parkinson's. At the molecular level, neurotoxicity is thought to originate from toxic gains of function in multimeric aggregates of proteins that are otherwise predominantly monomeric and disordered, fluctuating between a very large number of structurally dissimilar states on nano- and microsecond timescales. These proteins, termed Intrinsically Disordered Proteins (IDPs), are notoriously difficult to probe using traditional biophysical techniques. In order to obtain structural information pertaining to the aggregation of IDPs, it is often necessary to develop computational and modeling tools, both to leverage the full extent of the experimental data, and to generate testable predictions for future experiments. In this thesis, I present three separate computational studies studying the formation of multimeric aggregates in IDPs, spanning different aspects of the aggregation process, from early nucleation events to fibril elongation. In the first study, I present a conformational ensemble of α -synuclein, the culprit protein of Parkinson's disease, constructed using a Variational Bayesian Weighting algorithm in combination with NMR data collected by our collaborators. We find that the data fit a description in which the protein predominantly exists as a disordered monomer but contains small quantities of multimeric states containing both helical and strand-rich conformations. In the second study, I focus on the process of amyloid fibril elongation in the Amyloid- β (A β) peptide of Alzheimer's disease. I compute the free energy surface associated with the fibril elongation reaction, and find that elongation of both A β 40 and A β 42 experimental fibril structures occurs on a downhill free energy pathway, proceeding via an obligate, fibril-associated hairpin intermediate. The fibril-associated hairpin is significantly more stable (relative to the fibrillar, elongated state) in A β 42 compared with A β 40, suggesting a potential clinical target of interest. Finally, I present lengthy, all-atom molecular simulations that suggest that nucleation of the minimum aggregating fragment of α -synuclein proceeds via a helical intermediate, requiring a structural conversion into a strand-rich nucleating species via a stochastic process of individual helices unfolding and self-associating via backbone hydrogen bonds.

Thesis Supervisor: Collin M. Stultz

Title: Professor of Electrical Engineering and Computer Science, and Professor of Health Sciences and Technology

Acknowledgements

I am indebted to many people, only a subset of who are mentioned here. First of all, I must acknowledge my supervisor and mentor, Collin Stultz. I owe him my training and skills as a professional scientist. My relationship with him began as a visiting Masters student from the UK in 2009, and since then, he has honed my ability to conduct research efficiently and independently. Over the years in his group, he has taught me a tremendous amount about physical chemistry and biophysics, but I feel this goes without saying. Through Collin's mentorship, I have come to understand that the ability to conduct research effectively includes but extends beyond the technical details of scientific analysis. He has spent countless hours teaching me how write and present scientific material clearly and concisely, to navigate the emotional tribulations of the peer-review system, to mentor other students effectively and to foster collaborations with colleagues and other scientists. I now feel confident in my ability to conduct myself both as a scientist and as a professional in my future endeavors, and for this I am eternally thankful. I am also deeply grateful for his friendship, loyalty and sense of humor. I hope that the many great discussions we have had will continue long into the future.

In more ways than one, this thesis belongs to my parents, Sylvie and Francis, who have been an unwavering source of encouragement and support throughout my life. They transmitted to me a healthy disrespect for authority in conjunction with an appreciation and respect for learning, both of which have motivated my desire to pursue scientific research. My mother taught me the humility to free myself from the shackles of intellectual elitism that can sometimes come with an advanced education, and the open-mindedness to consider the limitations of logic and rational thought. My father taught me the values of scholarship and intellectual rigor from a very young age. Suffice it to say that I would not have chosen to pursue these values had I not been shown the way. I must also thank my sisters, Céline and Emma, who always kept my ego in check and whom I could always trust to have my best interest at heart. In a human society where incentives govern much of behavior and self-interest runs rampant, the value of unconditional fraternity is beyond measure.

I am also grateful to the members of my thesis committee, Professors Susan Lindquist, Bernhardt Trout, and Mark Bathe, for their insightful comments and guidance throughout the process that led to the content of this dissertation. In addition, I was very fortunate to enjoy a very fruitful and stimulating collaboration with Professor Tim Lu and Dr. Chao Zhong at MIT. Our work together opened my eyes to the potential of amyloids as a technological solution as well as a subject of scientific interest, and I am very appreciative for this insight.

Thanks is due to my labmates, past and present, whom I name without titles and who have been integral both to my education and to my mental health throughout graduate school: Paul Nerenberg, Yun Liu, Sarah Bowman, Linder Candido da Silva, Virginia

Burger and Molly Schmidt. In particular, I owe special thanks to Orly Ullman and Charles Fisher, with whom I had many valuable discussions, both on personal and scientific levels. Like trench mates, we witnessed each other's highs and lows, and explored the dirty and sometimes painful reality that lies behind scientific research, together.

I must also thank my friends and extended family, which through countless discussions have informed my vision of the world and the place of science in society. Jon, Ben, Chris, Ibrahim and Karim have given me the gift of humor and levity, without which it would have been difficult to survive the harsh realities of graduate school and scientific research. Ameer, Jeremy, Ossi and Thomas have motivated my sense of social responsibility and my desire to keep ethics and morality at the forefront of my decision-making process, while retaining a sense of wonder so critical to enjoying both life and science. Oguzhan and Giancarlo, my roommates during the first three years of graduate school with whom I shared countless adventures, were and always will always be a central part of the narrative of my Ph.D. years, in addition to my CSB 2010 classmates (Manoshi, Leyla, Mimi, Nate and Arshed), all of whom are in my good books. My cousins J and Kick have always expressed a faith in my abilities, despite us being half a world apart most of the time, and this faith has infused my self-doubt in such a manner as to make it palatable.

Most of all, I thank my wife and favorite person, Anu. I owe the smile on my face, and all that it reveals, to you.

Table of Contents

Introduction	7
Introduction to Intrinsically Disordered Proteins	7
Folded proteins versus IDPs	10
Experimental studies of IDP “structure”	15
Computational methods for describing IDP ensembles	17
Aggregation and neurodegeneration	21
Aβ mutations and aggregates	22
Aβ oligomers	24
Aβ fibrils	26
Transition between monomer and aggregate	27
Insight into Aβ through computation	29
Conclusions	30
The dynamic structure of alpha-synuclein multimers	34
Abstract	34
Introduction	35
Materials and Methods	36
Generation of seed structures	36
Generation of alpha-synuclein structural library	37
Generation of the ensemble and calculation of confidence intervals	38
Secondary structure assignments	41
Solvent accessibility calculations	41
NMR studies	41
Results and Discussion	44
Conclusions	52
The Mechanism of Amyloid-β Fibril Elongation	59
Abstract	59
Introduction	60
Materials and Methods	61
Model system	61
Reaction coordinates and umbrella sampling	61
Results	66
Discussion	82

All-atom simulations suggest that nucleation of an amyloidogenic peptide proceeds through a helical oligomeric intermediate	89
Abstract	89
Introduction	90
Results and Discussion	91
Materials and Methods	97
Conclusions and future directions	98
Appendix	102
A1 - Definition of f_β, a reaction coordinate quantifying strand content	102
A2 - Derivation of molecular dynamics forces arising from introducing umbrella potentials in f_β	104
A3 - Implementation in CHARMM	110
A4 - Theoretical foundations of umbrella sampling	111
Bibliography	114

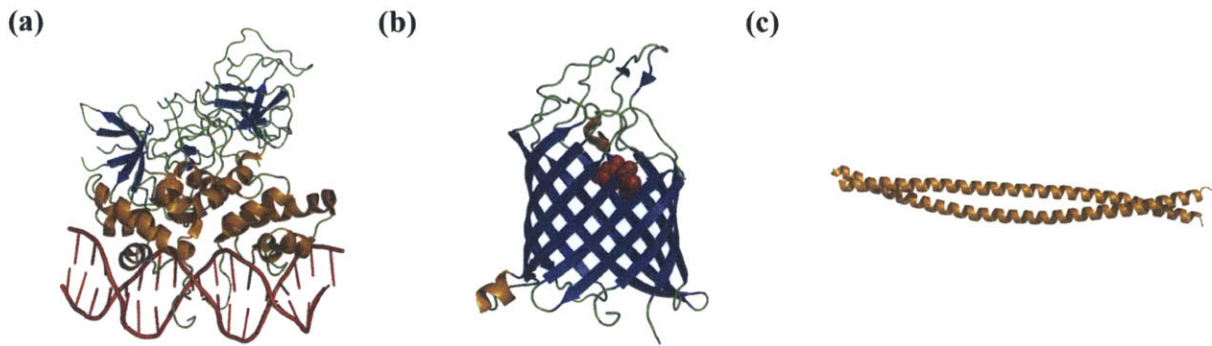
Introduction

Much of the work presented in this chapter was included in a review published in *Polymers*, Volume 6 (10), pp 2684-2719, October 23rd, 2014.

Introduction to Intrinsically Disordered Proteins

Proteins are fascinating heteropolymers that play essential roles in virtually all-biological processes. Their vast importance in biochemistry and medicine explains why a great deal of effort has been directed at understanding their properties and function. Traditionally, proteins have been understood to have a well-defined three-dimensional structure that is inextricably linked to their function. Indeed, knowledge of the structure of a protein provides a great deal of information about that protein's function (Fig. 1) (1, 2). The importance of protein structure is underscored by the fact that amino acid mutations in the primary sequence which destabilize the structure often result in disease

(3). The paradigm that a protein's structure determines its function has guided our understanding of proteins for decades.



*Figure 1. Examples of the relationship between protein structure and function. (a) Crystal structure of the Lambda-phage repressor (PDB ID 3BDN), which binds to its target DNA sequence (red) with high specificity. It achieves this with a helix-turn-helix motif (shown in orange) that can make sequence-specific contacts through the grooves in the DNA double-helix. (b) Crystal structure of Tsx (PDB ID 1TLZ), a nucleoside transporter protein, that transports nucleosides such as uridine (here shown in red) across the outer membrane of *E. coli*. It creates a pore in the membrane through a β -barrel motif, where the width of the cleft formed by the barrel, along with the individual side-chains that point into the cleft, determine what is allowed to travel across the membrane. (c) Crystal structure of keratin (3TNU), a fibrous structural protein whose toughness can be attributed to the helical coiled-coil structure it adopts in its fibers.*

Although proteins are often depicted as having static three-dimensional structures, thermal fluctuations at body temperature enable them to sample different conformations throughout their biological lifetime (4). Protein motions range from fast (\sim picoseconds) small amplitude (\sim Angstroms) fluctuations, to relatively slow (microseconds to seconds or longer), large scale motions that involve domain motions and/or folding (5). In

general, all of these motions enable proteins to perform their prescribed functions. Given the essential role that protein motion plays in biology, discussions about protein structure should ideally revolve around the structural ensemble of thermally accessible states that a given protein can adopt (6).

For a number of proteins, the structural ensemble consisting of its thermally accessible states contains structures that have only relatively small deviations from the ensemble average structure. In general, such proteins are categorized as being “folded”, and for these proteins, structures determined by experimental methods such as X-ray crystallography correspond to the ensemble-averaged structures. Since the folded ensemble contains structures that have only small deviations from the ensemble average structure, the ensemble average itself captures many important features of the protein’s structure, and many insights into a protein’s function can be garnered from this ensemble average structure (Fig. 1). By contrast, proteins within the more recently characterized class of intrinsically disordered proteins (IDPs) sample dissimilar conformations during their biological lifetime, and therefore the corresponding structural ensembles are heterogeneous. Given the vast number of structural states that are accessible to a disordered protein, the ensemble averaged structure for an IDP is typically not representative of any structure in the ensemble itself and therefore has little utility for understanding that protein’s function.

IDPs are quite prevalent in biology, despite having only been discovered in the last thirty years. It has been estimated that 25% of proteins encoded in the human genome are completely disordered and that 40% contain an intrinsically disordered region of at least 30 amino acids in length (7). These proteins have been found to play essential roles in many pathological processes. For example, aggregates of the IDP α -synuclein can be found in the brains of patients with Parkinson’s disease, and these aggregates have been linked to synaptic dysfunction in dopaminergic neurons (8). Huntington’s disease, another IDP associated neurodegenerative disease, is traceable to aggregation of the IDP Huntingtin protein, which contains glutamine repeats in its amino acid sequence (9-12). In the case of Alzheimer’s disease, aggregation of the IDPs Amyloid- β peptide ($A\beta$) and

tau protein are pathological hallmarks of Alzheimer's disease (13-15) (16, 17). In addition to diseases related to aggregation of IDPs, many diseases are caused by errors in signaling pathways. Mutations in IDPs involved in regulation of the cell cycle can disrupt gene regulation and cell signaling, mechanisms that are implicated in oncogenesis (18). Tumor suppressor p53 is a largely disordered protein, which functions in cell cycle regulation. Deactivating mutations of p53 can facilitate uncontrolled cell division and oncogenesis; e.g. mutations in the p53 are found in over 50% of cancers (19), including tumors of the colon, lung, esophagus, breast, liver, and brain (20).

While the importance of IDPs in human biology is not under question, their inherent structural heterogeneity makes them particularly challenging to study. In what follows, we first review protein structure in general, focusing on important differences between folded proteins and disordered proteins. We then introduce computational methods for studying intrinsically disordered proteins, and discuss examples of where these methods have been and could be applied to increase understanding of a specific IDP, A β , the aggregation of which will be considered in a later chapter.

Folded proteins versus IDPs

Proteins are heteropolymers consisting of covalent linkages between consecutive amino acids monomers, forming a chain. The amino acid sequence of a protein, termed its primary structure, confers chemical properties to the protein through the unique properties of the 20 different amino acids. For traditional “folded” proteins, this chain tends to fold into a unique structure. A central dogma of biochemistry is that a protein’s amino acid sequence determines its structure, which in turn determines its function (1, 21). While this paradigm is à propos for folded proteins, it is too simplistic to describe the vast array of experimental observations that have been made over the past few decades. Unlike folded proteins, intrinsically disordered proteins sample a variety of structurally dissimilar states during their biological lifetime, and therefore cannot be adequately described by a single well-defined structure (22).

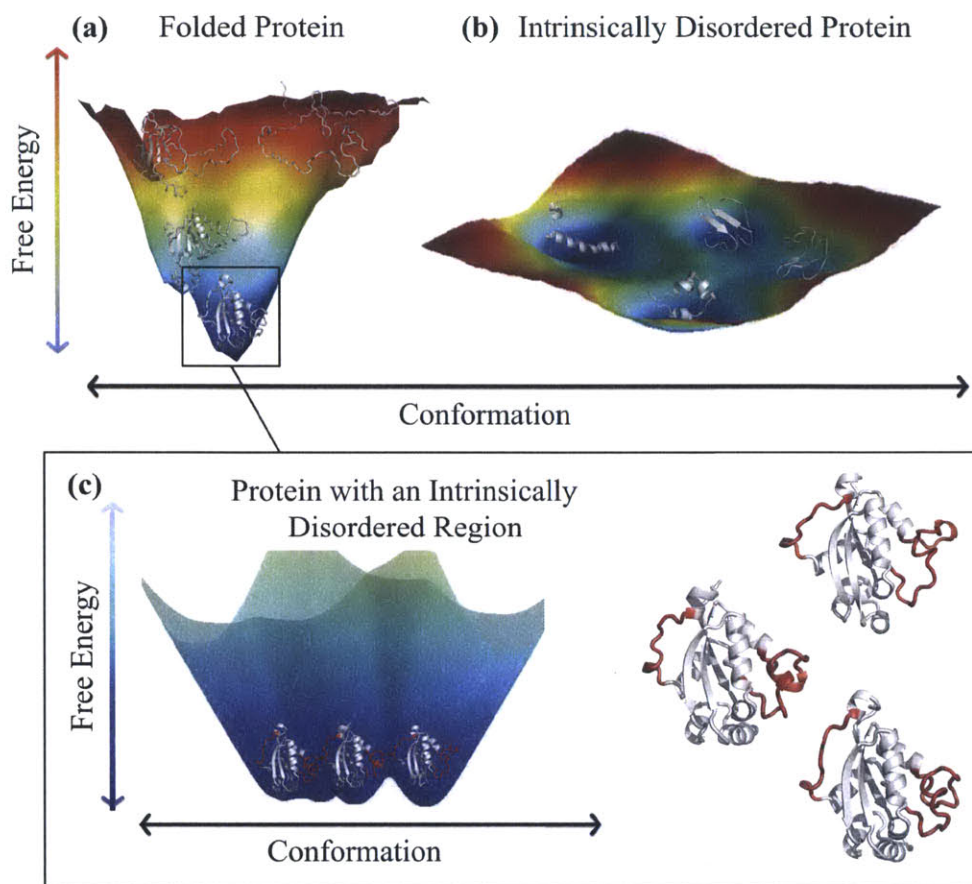


Figure 2. Schematic of energy landscapes for (a) a structured protein (human nucleoside diphosphate kinase (NDPK), PDB ID: 1nsk) and (b) an intrinsically disordered protein (CcdA C-terminal, PDB ID: 3tcj.). (c) Close-up of the minimal free energy well for NDPK, where IDRs are shown in red and structured regions are shown in white. Lower free energy (dark blue) represents more probable conformations The IDR structures are shown again enlarged to the right for better visualization.

The difference between folded proteins and disordered proteins can be understood based on an analysis of their free energy landscapes (Fig. 2). Folded proteins have a “funnel-shaped” global free energy minimum, where the lowest energy state corresponds to the

native structure (23) (24), and the width of the unique global energy minimum determines the conformational entropy of the native state (Fig. 2A). By contrast, disordered proteins have multiple local energy minima separated by small barriers (Fig. 2B). Transitions between the different local energy minima occur quickly and often, leading to an ensemble consisting of a vast number of structurally dissimilar states, which have approximately equal energies. Thus, a comprehensive characterization of an IDP consists of an ensemble of states and the transition rates between them (25). In practice, knowledge of the transition rates between conformers in an IDP ensemble is very difficult to capture, both experimentally and computationally. Consequently, in practice, studies of IDPs have focused on modeling the thermodynamically accessible states alone. As we outline below, while this represents an incomplete picture of these proteins, a great deal of information and insight has arisen from such studies.

While the above distinction between folded and disordered protein landscapes is instructive, it misses many of the nuances associated with discussions of protein structure. As we have alluded to above, all proteins sample a variety of different structures during their biological lifetime. Thermal fluctuations cause both folded and disordered proteins to sample a variety of states at temperature above 0K. In this regard, we note that even proteins considered to be folded (and whose structures have been solved via x-ray crystallography), often contain intrinsically disordered regions (IDRs), which lack a stable tertiary structure (26). This means that the energy minimum of a folded protein with an IDR is actually not smooth, but is actually a rough surface with many smaller minima corresponding to different states sampled by the IDR within the native state (Fig. 2c). Typical representations of folded and disordered proteins attempt to capture these inherent differences between the ensemble of states in a minimalist, yet informative manner. Folded proteins are often depicted as a single ensemble average structure, while disordered proteins are often represented by an alignment (or overlay) of energetically favorable, yet structurally dissimilar states (Fig. 3).

Disorder imparts a number of properties to IDPs that would be difficult for folded proteins to realize. For example, the structural heterogeneity of IDPs (and IDRs) confers

an ability to be promiscuous in their choice of binding partners (27, 28). This property explains why IDPs are frequently found to be hubs in protein interaction networks and are specifically associated with signaling networks (27, 29). In fact, almost 70% of signaling proteins are predicted to be intrinsically disordered (18). The largely disordered tumor suppressor p53, for example, is an important signaling hub, binding hundreds of proteins (29). An additional strength of IDPs in signaling networks is their fast production and degradation due to lack of stable structure, allowing them to be quickly activated or deactivated in response to changing cellular environments (30). Outside of signaling, some structural features are enabled directly through the flexibility of IDPs, such as the elastic properties of elastin (31).

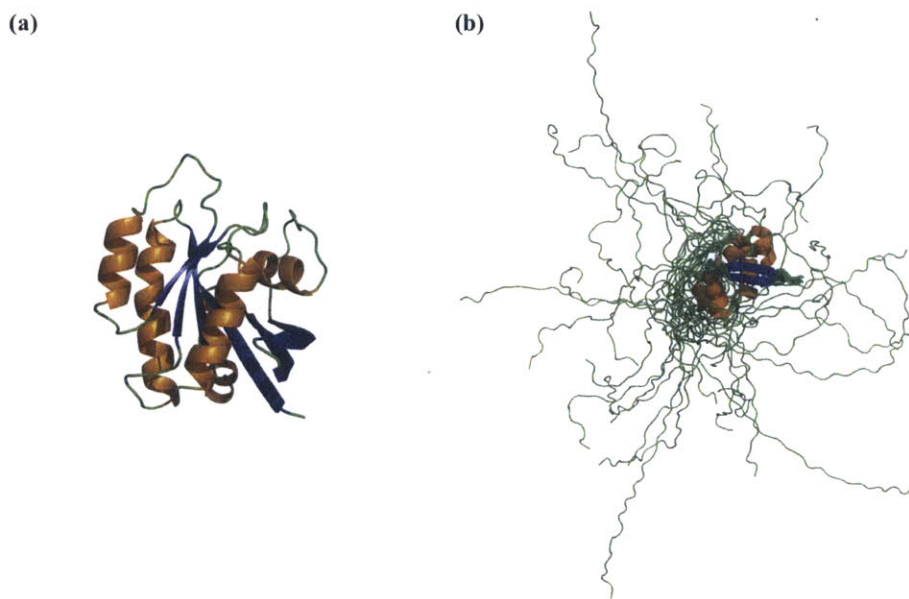


Figure 3. Varied degree of order in proteins. (a) Crystal structure of the protein H-Ras, solved in complex with GTPase-Activating Protein (not shown; PDB ID 3K9L). H-Ras is a folded protein containing a number of unstructured loops (shown in green) that have well-defined B-factors. For example, in the top loop, which is composed of residues 117-126, the backbone atoms have an average B-factor of 44.1\AA^2 , which suggests the loop is only somewhat flexible (compared to an average of

21.2Å² across the entire protein). These loops are unstructured yet they are ordered in the sense that they have well defined three-dimensional coordinates. Structured and ordered regions of the protein are shown in orange and blue according to their secondary structure. (b) NMR ensemble of a CcdA dimer (PDB ID 2H3A), a protein with both a folded region and an IDR. The intrinsically disordered C-terminal tail (shown in green) populates a large number of structurally dissimilar states. Each of the potential structures is depicted as distinct backbone traces in green, and the folded regions are shown in orange/blue according to secondary structure.

IDPs commonly obtain a folded structure upon binding their partners. Whether folding occurs before, during, or after contacting the partner is an oft-studied question, due to its implication for design of molecules to inhibit or stabilize IDP conformations. The conformational selection hypothesis proposes that IDPs fluctuate through their bound conformations while in the unbound state, and their partners selectively bind when the IDP is in the correct binding conformation (32). Alternatively, the induced fit hypothesis proposes that IDPs first make low-affinity, non-specific contacts with their partners, and then fold as they bind (33). Fly-casting, a related supposition that expands on this principle, states that extended IDP conformations results in a relatively large capture radius that accelerates the formation rate of these initial, low-affinity contacts. This provides a kinetic advantage for binding relatively to other structured proteins (34, 35). According to this hypothesis, the IDP folds into its bound conformation after the initial weak complex formation. While these hypotheses provide useful models for considering the formation of a protein complex involving an IDP, it is likely that the extent to which a binding event involves conformational selection, induced fit and/or fly-casting depends on the system in question.

Experimental studies of IDP “structure”

The ensemble average structure of a folded protein is usually determined using X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy (via the measurement of distance constraints between heavy atoms) (36). These methods, however, cannot be used to obtain a comprehensive picture of the structural ensemble of IDPs, for the reasons mentioned above. Techniques such as hydrogen-deuterium exchange NMR experiments aimed at probing the degree of solvent exposure of different regions of a protein’s sequence and useful in discerning loop regions and exposed surfaces are not applicable to IDPs since the majority of the protein is frequently exposed to solvent and the signal would be saturated.

Instead, lower resolution experimental methods can be used to find boundaries and distributions of measurable variables across the ensemble of conformational states sampled by the IDP, providing some measure of the underlying heterogeneity.

Insights into aspects of an IDP ensemble are typically obtained using a number of experimental techniques. Two useful methods are secondary chemical shifts determination and the measurement of paramagnetic relaxation enhancement (PRE). Secondary chemical shifts, measured with NMR, quantify the deviation between measured chemical shifts and random coil chemical shifts for each residue, providing information about secondary structure propensities in IDPs (37). It is important to note that since IDPs typically fluctuate between dissimilar conformations on a time scale that is fast relative to the experimental time scale, the measured chemical shifts at each residue are ensemble averages (38). NMR PREs measure long-range (up to 25 Angstrom) residual contacts within a protein by tagging a specific amino acid with a paramagnetic probe, thereby affecting the relaxation properties of nearby nuclei (38, 39).

NMR measurements of the Nuclear Overhauser Effect (NOE) can also provide short range distance constraints between different nuclei in a structure (38) (36). However, given the relatively large size of many IDPs and the conformational heterogeneity of their ensembles, NOEs between residues in the primary sequence are typically not observed in IDPs; i.e., on average nuclei from different residues are typically separated

by more than 5 angstroms (the typical limit for observing an NOE between nuclei) (36, 38, 40). Thus, while NOEs can be used to form distance constraints between residues for determining structure of folded proteins, these are typically not suitable for IDPs (41). Residual dipolar coupling (RDC) measurements provide long-range information about the protein's structure by measuring a partial alignment of the protein with respect to an external magnetic field. The protein of interest is typically embedded in an alignment medium that reduces the effects of molecular tumbling, after which the dipolar couplings are measured. RDCs encode information about the overall size of the molecule, and ¹H-¹⁵N amide RDCs, to some extent, encode information about secondary structure propensity (38).

Small angle X-ray scattering (SAXs) experiments provide information about the overall shape and size of molecules (42). Although these data, again, correspond to ensemble average information, when combined with structural models, SAXs profiles can provide important information that can be used to validate and refine models describing the thermodynamically accessible states of the IDP of interest. Recently, high speed atomic force microscopy (HS-AFM) has allowed visualization of the topography of proteins at nanometer resolution through a time-series of topographic images with a frame rate of more than ten frames per second (43). HS-AFM does not require labeling or staining of the molecule, but forms a topographic image of an entire system residing on a surface in a solution with minimal perturbation to the molecule in near physiological conditions (44). In studies of the 1767 residue heterodimeric protein FACT, which contains two major IDRs consisting of approximately 200 residues each, a frame-rate of 5 – 12.5 frames per second was sufficient to visualize changes in the IDRs' surface over time (45). While a higher frame rate would be necessary to visualize transitions between conformations or instantaneous snap-shots of molecules, these data can be used as bounds on models of IDPs, for example, in the form of distributions of radii of gyration. HS-AFM was additionally used to visualize formation of amyloid fibrils in amyloid-prone fragments of Lithosthatine (46). Again, while higher temporal resolution would be necessary to observe topographic changes resulting from the molecular processes

involved in formation of fibrils, these consecutive “snapshots” provide insights into the fibrillization process.

Computational methods for describing IDP ensembles

Molecular simulations can complement experimental methods, yielding structural models for the dominant thermodynamically accessible states of IDPs (47). While experiment usually provides ensemble-averaged information, molecular simulations provide atomistic, time-resolved information that can clarify experimental observations and that can provide fodder for future experiments (22).

Molecular dynamics simulations, in particular, can generate trajectories for proteins using an underlying potential energy function, which is used to calculate the forces on each atom (and consequently the motion of each atom) in the protein (48, 49). The potential energy function includes terms describing the energy associated with bond lengths, bond angles, and torsion angles, as well as long range forces arising from the Coulombic energy and the van der Waals interactions. The parameters defining each of these terms are learned either empirically or from *ab initio* calculations (48, 50). Several issues arise when applying these methods to IDPs. First, most parameterized force fields were developed for folded proteins, and therefore it is an open question as to whether all of the available energy functions are generally applicable to IDPs. While some more specific force fields have been developed with IDPs in mind (and fruitfully applied), it is not clear how generally applicable these methods are (51-55). More importantly, the conformational heterogeneity of IDPs calls for extensive simulations to ensure that the relevant regions of conformational space have been adequately sampled. In general, this process is extremely demanding from a computational standpoint.

Another method for conformational sampling, attractive due to its relative computational efficiency, is the statistical coil model approach in which one samples from empirical potentials to quickly generate an ensemble of states (56). The computational advantage of this approach stems from the fact that structures are typically constructed by independently sampling individual residue backbone dihedral angle conformations for

each residue in the protein. In this regard the potentials used are much simpler than molecular dynamics potentials and usually seek to reproduce coarse-grained behaviors, such as empirical backbone dihedral angle distributions for each residue from the Protein Databank (PDB) (56-58). Like molecular dynamics potentials, however, the empirical potentials used in statistical coil-based approaches are usually trained on conformational propensities of natively folded proteins; e.g., the backbone dihedral angles of residues designated as coil (e.g., regions not in strand or helical conformations) in the PDB. User-defined restraints can be included, such as done with the Flexible Mecanno tool (57), to adapt the potential to the particular peptide in question.

While there is much merit in these approaches, generating an accurate structural ensemble using these methods alone is not tractable for systems of even modest size. Computational tools may therefore have their greatest utility when used in conjunction with experimental data. For example, experimental observables can be used to restrain molecular simulations to obtain ensembles that have calculated observables that agree with the corresponding experimental values (59). Such ensemble-restrained simulations have been used to obtain conformational ensembles of alpha-synuclein by restraining molecular dynamics simulations with paramagnetic relaxation enhancement (PRE) measurements, which provide information about the long-range interatomic distances in the protein (60). These studies find that alpha-synuclein populates an ensemble of states that have smaller hydrodynamic radii than random coils, suggesting some degree of residual structure driven by interactions between the charged C-terminus and the hydrophobic central region of the protein (60). Other approaches first generate candidates for the thermally accessible states of the protein using an empirical potential energy function and then compare calculated ensemble averages from the molecular models to corresponding experimentally determined ensemble averages. Correct models have calculated averages that agree with experiment (61). These models and their associated experimental data can be deposited in an openly accessible database termed pE-DB (62). One example of such an approach is ENSEMBLE, which takes as input a set of conformations and experimental data, and prunes this large set of conformations to a

smaller set. Each structure is assigned a weight such that their ensemble average measurements agree with the data, and structures that do not contribute to fitting the experimental data are thrown out (63). Another approach for creating an ensemble that agrees with experimental measurements involves generating structures using a statistical coil-like model (*Flexible-Meccano*), a subset of which are selected for the agreement between their backbone dihedral angles and NMR chemical shifts. The process is then iterated until no further improvement in the agreement between chemical shifts and backbone dihedral angles can be obtained (64).

It is important to note that since experimental observables typically correspond to ensemble averages, it is not clear how to combine experiment with the results of computational models to arrive at an unfolded ensemble. While the problem of generating an ensemble that agrees with experiment is mathematically well defined, it has the uncomfortable consequence that experimental data collected on IDPs are inherently degenerate. More specifically, the number of experimental restraints one can obtain from any given experiment pales in comparison to the number of degrees of freedom associated with even the smallest IDP. In other words, one can generate many mutually exclusive structural ensembles that have ensemble averages that agree with any given set of experimental data (61, 65-67).

Several methods have been developed to deal with the degeneracy issue. In the most straightforward approach, one generates a number of different ensembles for an IDP that all agree with experiment. Structural features that are in common to all of the ensembles are interpreted as being those that are most likely to be “true”; i.e., while one cannot unambiguously determine which ensemble is correct, features that are common to all of the ensembles are likely to be legitimate (66, 68). A second method bases the choice of ensemble on a maximum entropy or, equivalently, a minimal information approach (69, 70) (71). The general principle ensures that the ensemble 1) yields calculated observables that agree with experiment; and 2) is as similar as possible to some pre-defined “prior” probability distribution. For example, if the prior distribution is given by the potential energy of the potential conformers, then the method yields an

ensemble that agrees with experiment and that minimally differs from what the potential energy surface says are favorable conformations.

Another method that explicitly tackles the issue of degeneracy is Bayesian Weighting (BW) (65, 72). The BW method consists of constructing coarse-grained conformational ensembles, defined as a finite set of representative states and an associated vector of weights, , which specifies the relative stabilities of each structure in the ensemble. The method begins by first generating a set of structures, either through a statistical coil model or by sampling from a molecular dynamics potential energy function. Predicted experimental measurements for each of these structures are then obtained using a variety of available algorithms (e.g. SHIFTX for NMR chemical shifts (73)). Using a Bayesian formalism, a posterior distribution over all possible weights for each structure is then computed by maximizing the agreement between the conformational ensemble and the experimental data. The strength of this approach lies in the fact that it accounts for both uncertainty associated with the experimental measurements (i.e. measurement error) and uncertainty in the algorithms used to predict experimental data from a given structure (i.e. prediction error) when generating the posterior distribution. Furthermore, it provides a quantitative estimate of the uncertainty in the underlying ensemble in the form of an uncertainty parameter, which takes a value between 0 and 1 and represents the extent to which one can assign weights to the structures in differently to agree with the data (65). This uncertainty parameter was found to correlate well with the error between reference ensembles and their corresponding constructed BW ensemble (65). Thus, the BW formalism allows the user to use quantitative experimental measurements, such as NMR or SAXS data, to construct conformational ensembles that include some measure of their statistical uncertainty. Given the highly degenerate nature of the data, it is helpful to construct one's structural library around a particularly quantity of interest, such as secondary structure content, and select representatives such that they cover the full range of possible values (61, 74).

To illustrate how computational tools can be used to provide information on the relationship between IDPs and disease, in the remaining sections we focus on A β , the

aggregation of which is linked to Alzheimer's disease. We discuss how the computational tools mentioned above can aid in the process of garnering detailed structural insights into the disease process, which can in turn be applied to the rational design of novel compounds aimed at combating disease.

Aggregation and neurodegeneration

Common to many neurodegenerative disease-related proteins is not only the disordered nature of the monomeric state, but also a tendency to self-associate to form a diverse range of aggregate states. The most conspicuous of these aggregates comes in the form of amyloid fibrils that can be isolated from brain tissue of patients who have died from one of these diseases, either as intra-neuronal depositions or tangles (in the case of α -synuclein, polyglutamine and tau) or as extra-cellular inclusions (in the case of $A\beta$) (75). An increasing body of evidence suggests that these fibrillar, amyloid structures are not the primary mediators of toxicity, but rather play secondary roles in the disease process, as either inert protein depositions at the end of the aggregation pathway or as secondary nucleation sites for the formation of smaller soluble aggregates (76). Instead, evidence suggests that lower molecular weight, soluble oligomeric aggregates are the primary mediators of toxicity in Alzheimer's and Parkinson's diseases (8, 15, 77-80). Whatever the precise disease causing species may be, it is clear that the aggregation process itself plays a pivotal role in the pathogenesis of these neurodegenerative disorders. A comprehensive understanding of the transition from a disordered state (an unfolded monomer) to an ordered, multimeric state (an oligomer or amyloid fibril), is therefore critical if one is to design novel therapeutics aimed at preventing or reversing this aggregation process (Fig. 4).

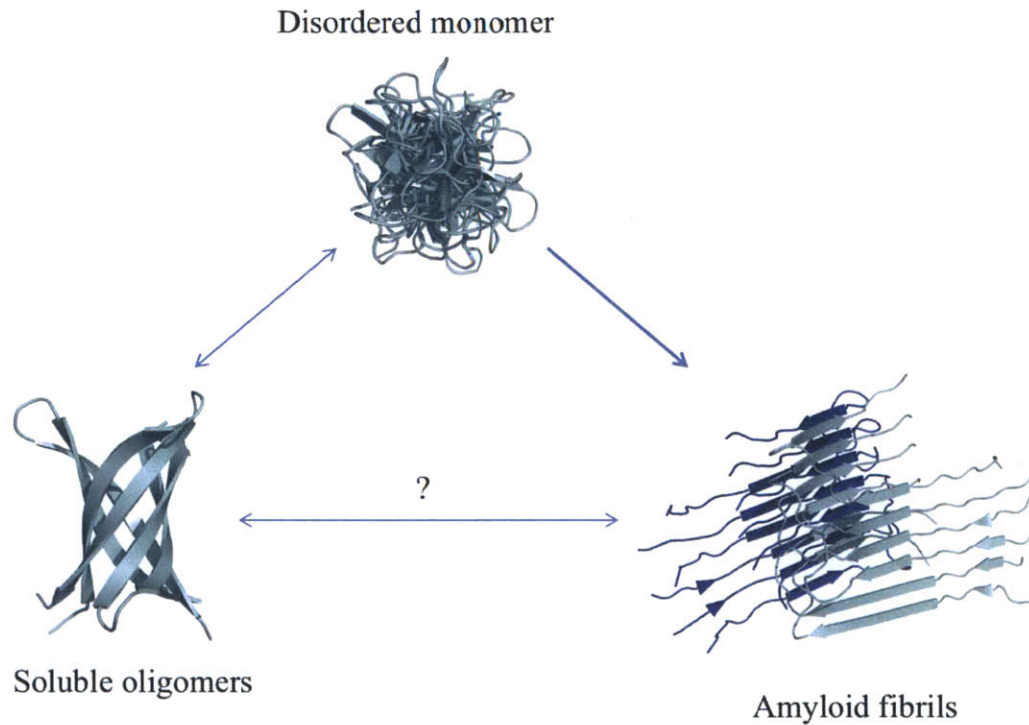


Figure 4. Schematic of the different “structures” of the A β peptide. Monomers can form fibrils, which are highly stable and rarely dissociate back into monomers, but can also form meta-stable, soluble oligomers. A hypothetical structure of a soluble oligomer is shown, which was constructed by threading the A β sequence to a published crystallographic structure of α -crystallin oligomers (81). A double-headed arrow between oligomers and fibrils is shown to illustrate a potential, but relatively unknown, interplay between the two species.

A β mutations and aggregates

Post-mortem examinations of the brains of patients suffering from Alzheimer’s disease (AD) have led to the identification of extracellular plaques in the cerebral cortex that test positive for the presence of a small, 4 kDa peptide called Amyloid β -protein (A β). A β was first purified from amyloid fibrils isolated from brain meninges in 1984 (82). It is the product of targeted proteolysis of the β -amyloid precursor protein (APP), a large single-transmembrane glycoprotein that is widely expressed in both neural and non-

neural cells (83). APP is first cleaved in the extra-cellular lumen by β -secretase to produce a membrane-bound C-terminal fragment, along with an extra-cellular N-terminal fragment that is secreted. The membrane-bound APP portion is then cleaved by γ -secretase to release the final A β peptide, which in APP is partially buried within the membrane (84). γ -secretase can cleave APP at multiple positions, resulting in A β peptides of different lengths. These peptides vary in the number of hydrophobic residues in their C-terminus, and as such have different aggregation propensities (85). Several mutations have been identified as being related to AD pathology. Some of these mutations are primarily found at, or directly flanking, the cleavage sites for the secretase enzymes, resulting in different distributions of cleavage products from the wild-type (86), while others are located within the central hydrophobic region of the cleaved A β sequence (87). For example, comparison of the carboxyl-terminal peptides produced from cleavage of wild-type versus mutant APP, the particular mutations of which have been linked to familial AD, showed an increase in the fraction of 'long' A β (particularly A β residues 1-42, or A β 42 for short) relative to A β 40 in the mutants (88). Studies of various lengths of A β show that longer A β fragments (A β 42 in particular) have an increased tendency to aggregate and form fibrils than the dominant form (A β 40) in wild-type cells (85).

NMR studies suggest that A β exists predominantly as a disordered monomer (16, 89). However, as previously mentioned for aggregating IDPs in general, the disease process in A β is associated with a transition from this disordered monomeric state to more ordered multimeric states. A β has been observed, *in vitro*, to form aggregates of varying molecular weight, spanning the range from small, low molecular weight soluble oligomers, through protofibrils (small assemblies of A β that nucleate the formation of larger amyloid fibrils), all the way to insoluble amyloid fibrils consisting of thousands of monomers in a highly repetitive configuration.

In the following sections, we first outline current knowledge of each aggregate state of A β , as well as open questions about each state and transitions between states. We then discuss how computation has addressed some of these questions.

A β oligomers

It is proposed that the pathogenesis stems from a toxic gain of function when these multimeric states are formed (78, 90, 91). A β appears to exist in a range of different oligomeric forms, presumably originating from disordered monomeric pools.

Characterization of oligomeric species of A β is particularly nebulous, compared to other A β species, due to their polymorphic nature. A β oligomers have been known to adopt a variety of molecular weights, morphologies, and secondary structure content (78, 80, 83, 92). Central questions surrounding the different oligomeric species are whether or not they constitute toxic entities, and whether their formation is on the pathway towards amyloid fibril formation, or occurs through independent pathways. Answering these questions is central to understanding the mechanistic basis behind the disease, and in addition might provide clues as to how these pathways could be manipulated to prevent or reverse the disease process.

The mechanistic basis for the neurotoxicity of oligomeric structures remains unclear (83). Early studies of A β suggest that it can form cylindrical, β -barrel type oligomers which resemble bacterial porins in electron micrographs (93). It is thought that such oligomers can create channels in the cell membrane, leading to Ca²⁺ dysregulation and disruption of the membrane's partitioning function (94). An analysis of HypF-N oligomers, which have similar properties to their A β counterparts, found that toxic oligomers produced an influx of extracellular Ca²⁺ into the cytosol, in contrast to non-toxic oligomers produced under different conditions, despite having the same morphological and tinctorial features (95). The same study found that the toxic forms differed in the packing of the hydrophobic interactions between adjacent monomers, suggesting that structural flexibility and hydrophobic exposure are critical determinants of an oligomer's toxicity (95).

There are very little data pertaining to the conformation of individual monomers in the toxic oligomers. The formation of soluble oligomers was not disrupted by stabilizing

monomeric A β in a β -hairpin state through the introduction of cysteine mutations in pairs of residues found to be in close contact in a solution NMR structure of the hairpin in complex with an Affibody, suggesting that these oligomeric species are composed of monomers in a similar hairpin state (96, 97). Amide-proton exchange NMR experiments have identified regions of the sequence that have the highest accessibility to the surrounding solvent when in a toxic oligomeric state. These regions are likely to correspond to turn conformations, and propose a configuration of strands arranged according to these turn regions (98). These findings are all consistent with the formation of cylindrical oligomers composed of individual β -hairpins or sheets, much like the crystallographic structure of cylindrin, an oligomeric form of alpha-crystallin fragments (81). Indeed, extrapolating from the structure of cylindrin, Laganowsky et al. propose a similar model of a trimeric A β oligomer (81). Such a structural arrangement differs fundamentally from a pre-fibrillar oligomer (e.g. a small protofibril) in that it cannot be extended naturally to include more monomers. This is because many of the hydrogen bond donors and acceptors of the polypeptide backbone that are involved in fibrillar, inter-molecular hydrogen bonds are bonded to each other in an intra-molecular fashion in the hairpin state (97). Thus, it is unlikely that these structures form the basis for further aggregation without undergoing some structural changes to adopt the cross- β arrangement of a prototypical amyloid structure.

By using the technique of photo-induced cross-linking of unmodified proteins (PICUP), it was found that aggregate-free samples of A β 40 contained monomers, dimers, trimers and tetramers in rapid equilibrium. In contrast, A β 42 preferentially forms pentameric and hexameric 'paranuclei' which assembled further into bead-like structures resembling protofibrils, arguing that the A β 42 assembly pathway involves the formation of distinct intermediates that gradually rearrange into protofibrils (99). Further studies combining mutational experiments with PICUP suggest that the side-chain of residue 41 is important for paranucleus formation and further self-association into larger oligomers, while the side-chain of residue 42 primarily impacts paranucleus self-association (100). A different study introducing the technique of ion mobility coupled with mass

spectrometry analyzed the *in vitro* oligomer size distributions for both A β 40 and A β 42 and found that they differed considerably, lending further evidence to the notion that A β 40 and A β 42 self-assemble along different pathways (101). *In silico*, coarse-grained simulations using a four-bead model which includes backbone hydrogen bonding, and residue-specific interactions due to effective hydrophathy and charge, found that A β 40 forms significantly more dimers than A β 42, while A β 42 forms more pentamers. Furthermore, they found that a turn centered around Gly-37-Gly-38 is formed in A β 42 and not in A β 40, and was found to be associated with initial contacts formed during monomer folding (102). A later study using the same simulation technique on Arctic mutants of A β 40 and A β 42 were used to derive size-distributions in agreement with prior experimental data, and showed that the A β 40 mutant was able to form paranuclei much like A β 42, although the mutations prevented aggregation into higher order oligomers in both isoforms (103). Using discrete molecular dynamics simulations of wild-type A β 40 and, Urbanc *et al.* found that the region D1-R5 is more disordered and exposed to solvent in A β 42 than A β 40, suggesting that the N-terminal region is involved in mediating toxicity (104).

A β fibrils

Histopathologic analyses of brain tissue derived from post-mortem examinations of patients that suffered from Alzheimer's disease reveal large inclusions in the neural tissue that are composed of large quantities of amyloid fibrils (105, 106). It has been suggested that a propensity to form stable amyloid structures under the right conditions is wide-spread across the proteome (107). These fibrillar structures are held together through intermolecular hydrogen bonds between the backbones of adjacent monomers arranged in β -strands perpendicular to the fibril axis, termed a cross- β structure (107-109). They are ordered and highly structured, insoluble in nature, and have well-defined and highly repetitive structural cores. Amyloids thus have proved to be somewhat more yielding to structure determination techniques (108, 110). Structural

models of A β fibrils derived from solid-state NMR restraints suggest a high degree of polymorphism in the different fibrillar structures. These models suggest that fibrils frequently contain more than one filament, such as the twofold and threefold symmetric fibrils of A β (108, 109, 111) which can be observed through scanning electron microscopy to be arranged in helical superstructures termed β -helices (112, 113). The solid-state NMR restraints used to create the twofold and threefold symmetric fibrils of A β were compatible with two mutually exclusive models for the relative height of anti-parallel β -strands within monomers in the fibril for both, termed positive and negative stagger (109). Extensive molecular simulations conducted on fibrils containing the two types of stagger found that only negative stagger fibrils formed the left-handed helical suprastructures observed by electron microscopy (112, 113). Initially, two competing quaternary structure contacts between the C-terminal strands of the two filaments were proposed based on molecular simulations: parallel and anti-parallel (114). Further solid-state NMR data indicated anti-parallel contacts between C-terminal strands (108). When simulated using coarse-grained molecular simulations, Fawzi *et al.* found that both models for the quaternary contacts were stable, but the anti-parallel model was more likely to elongate (115).

The N-terminal region of A β appears disordered even in the fibrillar state, with the remaining residues adopting the fibril core cross- β structure. This fibrillar conformation therefore suggests that, given the appropriate binding partner, there is a strong propensity for the formation of β -strands in the A β sequence.

Transition between monomer and aggregate

Very little is known about the structural basis of the transition from the disordered monomeric state to the ordered multimeric states. Based on our current understanding of the putative toxic oligomeric species, it is likely that the folding pathways that lead to the pathology associated with Alzheimer's are pathways involving the formation of β -strands (78, 83). A mechanism has been proposed involving the sampling of extended,

strand-based conformations of monomeric A β and alpha-synuclein that lead to exposure of the hydrophobic segments, which prefer to self-associate than to interact with the surrounding solvent (116, 117). Indeed, mutations that are associated with early-onset Parkinson's disease have been shown to decrease the rate at which the backbone of the alpha-synuclein protein changes its configuration, which would prolong the exposure of such segments (117). This type of mechanism could involve the formation of fibril-like, intermolecular hydrogen bonds between two colliding monomers with temporarily exposed backbones. The presence of a neighboring A β molecule in a particular conformation may alter the conformational landscape of the incoming protein, increasing its propensity to a particular β -strand state by an induced-fit type of mechanism, such that it would lead to the formation of oligomers and/or protofibrils which can then progress down the amyloid pathway. Indeed, 'seeding' an in vitro monomeric solution of A β with pre-formed amyloid fibrils causes these fibrils to extend readily (118). Moreover, fibrillar A β has proven to behave like a prion: when mice brains are inoculated with A β in a fibrillar form, rapid cell-cell transmission of the pathological species was observed (119, 120). This prion-like quality suggests that the presence of A β fibrils can alter the propensity of the monomer pool to adopt the fibrillar conformation. In contrast, currently available data for oligomeric A β suggest that oligomer-prone conformations may be sampled directly in the monomeric state (81, 96-98, 121). These suggest that aggregation could occur through conformational selection from the native monomeric ensemble, i.e. pre-formed states such as hairpins associate directly without major modification upon binding. Since A β 42 is known to form oligomers more readily than A β 40, it is therefore interesting to look for clues in the monomeric ensembles of each construct.

While the strand segments within each hairpin correspond to segments that are also in a strand conformation in the fibrillar state, the tertiary structural arrangement of these strands is different, since they are involved in intramolecular hydrogen bonds with either other (97), in contrast to the fibrillar conformations which contain no intramolecular contacts (108-111, 122). An oligomeric species composed of hairpin-type monomers

containing intramolecular hydrogen bonds would have to undergo significant structural rearrangements to form amyloid protofibrils, a process likely to involve a large kinetic barrier. For this reason, it seems unlikely that hairpin-based oligomeric species and protofibrillar oligomers are on the same folding pathway. However, monitoring the aggregation of a di-cysteine mutant of A β 40 *in vitro* by the selective binding of the latent fluorophore FAsH to oligomers and fibrils showed that A β 40 forms spherical oligomers that can slowly convert to amyloid fibrils through a nucleated conformational conversion mechanism (123). Furthermore, discrete molecular dynamics simulations of both A β 40 and A β 42 showed assembly of elongated protofibrils from spherical oligomers (103). These results are consistent with a number of studies having provided evidence for the formation of oligomers prior to the appearance of fibrils (99-101, 124). More recently, kinetic studies of A β 42 showed that the formation of toxic, soluble oligomers occurs as a secondary nucleation process, in which oligomers are formed in two phases: the first is in the absence of any amyloid aggregates, and the second in their presence (76). The second phase results in an increased rate of oligomer formation, and radiolabeling experiments confirmed that oligomers formed were derived from the monomeric pool of A β 42 rather than by breaking off fibrils directly. Thus, amyloid fibrils and toxic oligomers may form through distinct folding pathways, but the kinetics of oligomer formation is enhanced in the presence of fibrils. These data highlight the complex interplay between the monomeric, oligomeric and fibrillar pools of A β that is likely to underlie the disease state (Fig. 4).

Insight into A β through computation

Several studies have applied brute-force, unbiased molecular dynamics simulations of the A β peptide to explore the conformational preferences of the disordered monomer. One study, which totaled over 200 μ s of simulation time for each peptide, found that A β 40 and A β 42 have crudely similar characteristics, in that they can both adopt strand-based conformations, but that A β 42 has an increased propensity to form hairpins in its C-

terminus when compared to A β 40 (125). The conformational ensembles of the A β 40 and A β 42 monomers were constructed using BW with NMR data to learn the states sampled by each monomer (121). A set of structures $\{s^i\}$, generated through both REMD simulations of both full-length A β 42 and overlapping A β 42 peptide segments, used to construct both ensembles (with the last two residues of A β 42 truncated to form the A β 40 structure set), and weights $\{w^i\}$ were computed for both ensembles using their respective NMR data (121). Comparison of these two ensembles suggested a statistically significant, tenfold increase in the relative stability of a hairpin conformation in the A β 42 isoform versus its shorter counterpart, which provides a potential mechanism for its increased aggregation propensity (121) and correlates well with findings from unbiased molecular dynamics simulations of these two peptides (125). This finding is consistent with a conformational selection hypothesis involving hairpin structures (121). As discussed above for binding of the p53 termini to their interaction partners, evidence of the bound state in the unbound ensemble supports the role of conformational selection in binding, but does not explain the role of induced fit. Further studies probing the conformational landscape of A β in the presence of additional A β molecules could provide insight to the role of induced-fit in the formation of oligomers or protofibrils. Furthermore, computational studies could be employed to investigate the role of flexibility in toxic oligomers, as well as the different pathways to oligomer and fibril formation.

Conclusions

IDPs play a central role in many cellular processes, as their disordered nature provides them with the ability to bind many partners, thereby regulating many biochemical processes. Because of this central role, the malfunction of IDPs can disrupt proper cellular function and lead to disease. Unfortunately, their disordered nature, which makes them so relevant in cellular networks, also makes them difficult to study with

traditional experimental methods that were initially designed to study folded proteins. We discussed recent studies that have employed computational methods to analyze the conformational preferences and mechanisms of IDPs. We focused on A β , which is found in an aggregated state in the brains of patients who died of Alzheimer's disease. Understanding how A β transitions between disordered monomers and the different species mentioned above is a pre-requisite to controlling the early events of the Alzheimer's disease processes. We have shown that computational tools can provide some measure of leverage when analyzing quantitative, experimental structural data about the disordered state. This can be achieved by using empirical molecular mechanics force fields to understand the unfolded state of these polymers, as well as by computing a distribution for the ways in which one can weight a given set of structures with experimental data to generate a conformational ensemble, as in the Bayesian Weighting approach. Computational data are helpful in understanding the properties of the monomeric state and the mechanism of aggregation or abnormal signaling. Single-molecule experiments are showing promise in their ability to investigate the kinetics of conformational changes in a given monomer, which may lead to new insights into the aggregation process. Due to the highly ordered and structurally repetitive nature of amyloid fibrils, it has been possible for high resolution models of different fibrillar states to be developed. These results suggest that even the amyloid state is polymorphic and likely to be dependent on the nucleation species (108, 109, 111). The species that have proved most resistant to characterization unfortunately appear to be the most important: soluble oligomeric aggregates. We have discussed how current data suggest that hairpin-type conformations are present within the toxic oligomeric states of A β , thus distinguishing them from amyloid pathways due to the structural dissimilarity between hairpins and monomers in fibrillar conformations. Despite all of this, high resolution information about the transition from a flexible monomer to a folded, relatively rigid oligomer or fibril have proved elusive so far. Part of the difficulty may stem from the fact that monomers and oligomers are in fast exchange with one-another, as suggested

by data collected from multimeric alpha-synuclein, and computational studies could be targeted towards overcoming this obstacle.

One difficulty in characterizing IDPs stems from a lack of experimental and computational tools for studying folding events that occur on a timescale that is too fast to be probed with traditional experimental methods, and too slow to be tractable by traditional molecular simulations. A comprehensive understanding of this transition will therefore require improvements in the experimental methods available for structural characterization of short-lived intermediate states, coupled with a creative use of computational methods to obtain mechanistic insights into the transitions between these states.

The remainder of this thesis details three separate studies involving aggregated states of IDPs at various stages of the aggregation process. The first chapter investigates the existence of different types of oligomers in recombinant alpha-synuclein, an IDP involved in Parkinson's disease, by constructing what we believe is the first conformational ensemble of an IDP that contains multimeric states as well as monomeric. The second chapter proposes a molecular mechanism for the elongation of experimentally-derived models of A β amyloid fibrils. Finally, the third chapter performs an all-atom simulation of the early events of nucleation in the aggregation of an 11-residue alpha-synuclein fragment that is known to be disordered in the monomeric state, induces toxicity in cells and aggregates to form fibrils. All three chapters concern themselves with the formation of ordered aggregates in an otherwise monomeric IDP. The latter two chapters emphasize the characterization of the transition between a disordered monomer and a folded, ordered aggregate.

The dynamic structure of alpha-synuclein multimers

The work presented in this chapter was published in the *Journal of the American Chemical Society*, Volume 135 (10), pp 3865-3872, on February 11th, 2013. It represents the combined work of all co-authors on the paper.

Abstract

Alpha-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 alpha-synuclein can form beta-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 alpha-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 beta-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 alpha-synuclein when the protein concentration is high; thereby preventing non-membrane bound monomers from aggregating.

Introduction

Alpha-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(126). The most notable pathological characteristic of these diseases is the aggregation of alpha-synuclein into amyloid fibrils, which have significant beta-sheet secondary structure(105, 127). Although there is disagreement regarding whether the soluble oligomeric aggregates or insoluble aggregates are the most neurotoxic species, it is clear that alpha-synuclein self-association plays an integral role in neuronal dysfunction and death(8, 77, 128-130). 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 alpha-synuclein is notoriously difficult to study, earning it the moniker of 'chameleon' due to its tendency to adopt different conformations under different experimental conditions(131, 132). This has led to seemingly contradictory data about the dominant putative states in solution versus those under physiologic conditions(92, 133, 134). While it is clear that monomeric alpha-synuclein is an intrinsically disordered protein(135) in solution, recent data suggests that it can adopt a tetrameric state that has a relatively high helical content under physiologic conditions(92, 134, 136). By contrast, others have suggested that alpha-synuclein retains its monomeric disordered state *in cellulo*(133, 137).

Recently, NMR studies on an alpha-synuclein construct isolated from *E. coli*, which contains a 10 residue N-terminal extension, suggested that the protein can exist as a "dynamic tetramer"(134). 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 alpha-synuclein construct can adopt, we generated an atomistic model for alpha-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 $S = \{\bar{s}_i\}_{i=1}^n$, where \bar{s}_i is the Cartesian coordinates of structure i , and a corresponding set of weights $\bar{w} = \{w_i\}_{i=1}^n$, 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(134). Since we had previously constructed an ensemble for monomeric alpha-synuclein using NMR chemical shifts, RDCs and SAXS data(138), we used these structures to represent the disordered, monomeric fraction. Using NMR chemical shifts and NH RDCs obtained on an alpha-synuclein construct, which contains a 10 residue N-terminal extension, we determine the relative fractions of different multimeric forms within the ensemble.

Materials and Methods

Generation of seed structures

Our previous study on alpha-synuclein suggested that the monomeric, protein can sample amphipathic helices, which could in principle self-associate to form helical trimers and tetramers(138).

All simulations used a model of alpha-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 alpha-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)(139), 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) (140, 141). 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(134). 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 (H α -HN i , $i+3$) NOEs in ¹⁵N-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.

Generation of alpha-synuclein structural library

The conformational space of alpha-synuclein was sampled by subjecting the initial seed structures to replica exchange molecular dynamics (REMD) simulations(142). Each initial structure underwent REMD with the EEF1(143) implicit solvent model as implemented in the CHARMM(144) force field [ENREF_23](#). 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(65, 68, 138). 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(134), ensuring 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 alpha-synuclein (138) to yield our structural library $S = \{\vec{s}_i\}_{i=1}^{533}$ of 533 conformers.

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(72), which is a variational approximation to a Bayesian Weighting formalism used in the past(65, 138). 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:

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

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(72). Experimental observables, specifically C α , C β , N, H and H α chemical shifts from a previous work(134) in combination with backbone NH residual dipolar couplings (RDCs), were used (Supplemental Information Table S1). Predicted measurements for each conformer were generated using SHIFTX(145) for chemical shifts and PALES(146) 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)(72), 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 (72):

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

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 $g(\bar{w}|\bar{\alpha},S)$ and $f_{\bar{w}|\bar{M},S}(\bar{w}|\bar{m},S)$ is then minimized to find the optimal set of Dirichlet parameters, $\bar{\alpha}' = \{\alpha'_i\}_{i=1}^N$, 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 $\bar{w}^B = \{w_i^B\}$, which is the expected value of the vector of weights over the new approximate posterior distribution:

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

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

$$w_i^B = \frac{\alpha'_i}{\alpha'_0} \quad (4)$$

where $\alpha'_0 = \sum_i \alpha'_i$. The uncertainty parameter σ_{w^B} , called the posterior expected divergence, corresponds to the average distance from the Bayes weights over the entire space of weights:

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

where $\Omega^2(\bar{w}^B, \bar{w})$ is the Jensen-Shannon divergence, a metric which quantifies the distance between the vectors \bar{w}^B and \bar{w} (65).

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

$$\text{cov}(w_i, w_j) = \frac{\alpha'_i \alpha'_0 \delta_{ij} - \alpha'_i \alpha'_j}{\alpha_0'^2 (\alpha_0' + 1)} \quad (6)$$

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:

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

95% confidence intervals can then be computed using a Gaussian approximation from $CI = 1.54 \times 1.96 \times \sqrt{\text{var}(D)}$, 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(72).

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.

Secondary structure assignments

Secondary structure was assigned using DSSP(147). 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.

Solvent accessibility calculations

Solvent accessible surface area (SASA) was calculated for each conformation using CHARMM(144). Since only the backbone atoms N, H, C, Ca 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 alpha-synuclein backbone atoms when in a fully extended conformation.

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% D₂O. All NMR spectroscopy was performed on a Bruker Avance 800 NMR spectrometer operating at 800.13 MHz (¹H), 81.08 MHz (¹⁵N) and 201.19 MHz (¹³C) and equipped with a TCI cryoprobe and pulsed field gradients. Experiments used for sequential resonance assignments include three-dimensional (3D) experiments HNCA, HNCACB, ¹⁵N-HSQC TOCSY and ¹⁵N-HSQC NOESY. Quadrature detection was obtained in the ¹⁵N dimension of 3D experiments using sensitivity-enhanced gradient coherence selection(148), and in the ¹³C 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 (¹⁵N) were used. For ¹³C, 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 (149).

¹H-¹⁵N, ¹³C'-¹⁵N and ¹³C'-¹³Ca residual dipolar couplings (RDCs) were recorded for a ¹⁵N- and ¹³C-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.(150) 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 ²H water signal.

We used solution NMR to localize the transient formation of α -helices in α Syn. 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 (Ha-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 (Figure 1). 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.

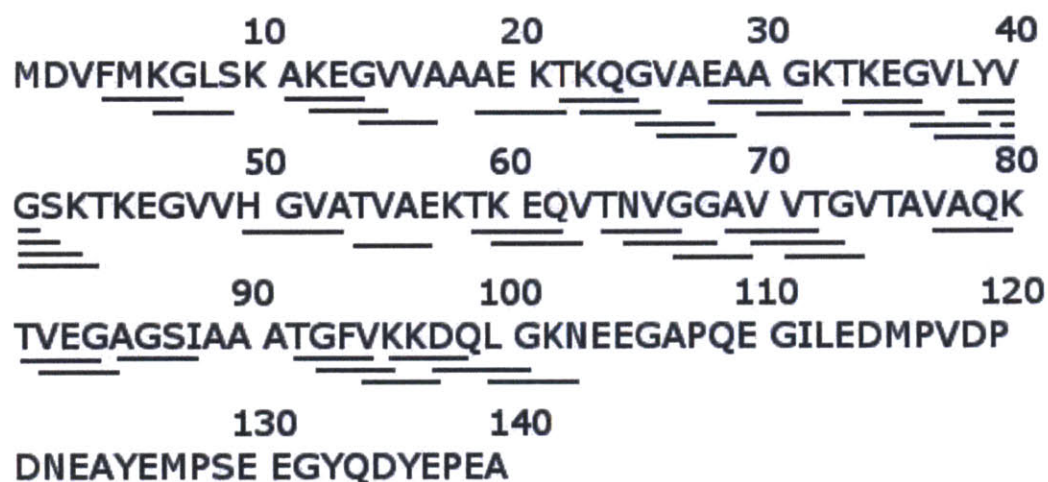


Figure 1. Regions of α Syn fractionally occupying helical structures as defined by i , $i+3$ Ha-HN NOEs. The experimentally determined NOEs used to construct the initial model are indicated by solid lines.

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(151). 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(152), and initial values for alignment tensors determined by singular value decomposition (SVD) using the program PALES(146). 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 (153). All NMR experiments were performed by Iva Perovic and Thomas Pochapsky. NMR data are available in the Supplementary Information of the published work (Table S1), available in print at <http://pubs.acs.org/doi/suppl/10.1021/ja310518p>.

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 alpha-synuclein suggested that the monomeric protein can sample amphipathic helices, which could in principle self-associate to form higher order structures (138). 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 (134). 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(142) (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 alpha-synuclein multimers had significant helical content(92, 134, 136). 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, $\bar{w} = \{w_i\}_{i=1}^n$, 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(61, 65). 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, $0 \leq \sigma_{\bar{w}^B} \leq 1$, which describes the spread of the posterior distribution – a metric that is akin to the standard deviation of a Gaussian distribution(65, 72). Our prior work suggests that the numeric value of $\sigma_{\bar{w}^B}$ is correlated with model correctness. When $\sigma_{\bar{w}^B} = 0$, we can be relatively certain that the model is correct. By contrast when $\sigma_{\bar{w}^B} = 1$, it is likely that the ensemble is far from the truth. Nevertheless, when $\sigma_{\bar{w}^B} \neq 0$, 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, $\bar{w}^B = \{w_i^B\}$, which corresponds to the expected value of the weights calculated from the posterior distribution, and the structural library $S = \{\bar{s}_i\}_{i=1}^n$. 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 2), the corresponding uncertainty parameter is

non-zero: $\sigma_{w^B} = 0.47$. Consequently, we express ensemble average values along with their corresponding 95% confidence intervals.

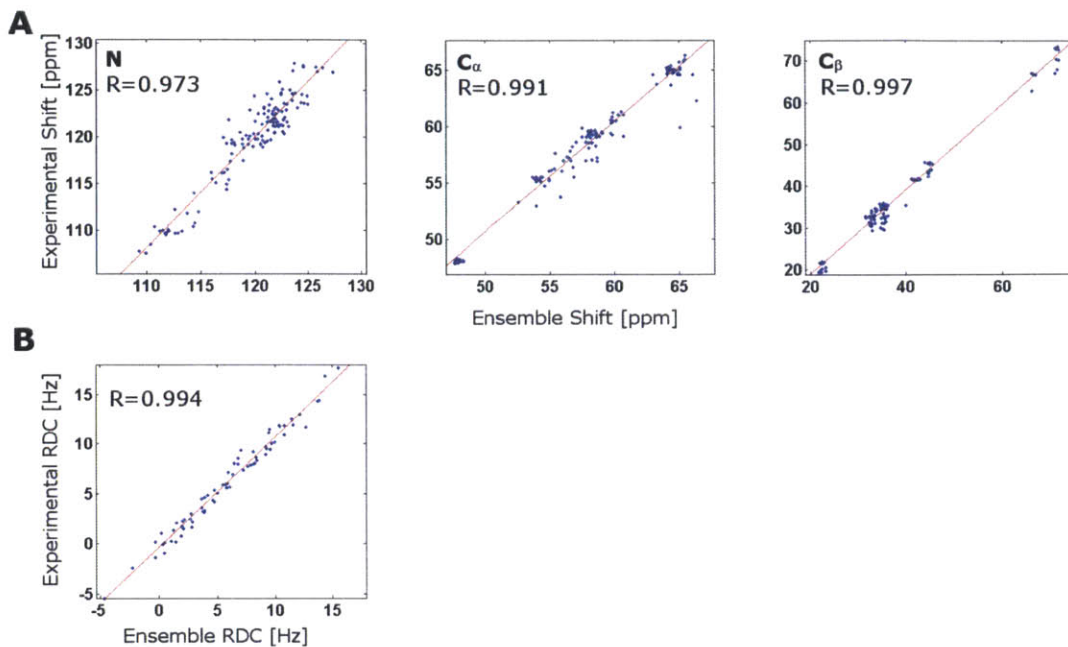


Figure 2. Calculated ensemble averages vs. experimental measurements. (A) N, Ca and C β chemical shifts; (B) N-H residual dipolar couplings. Correlation coefficients for each plot are explicitly shown.

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(138), 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 3.

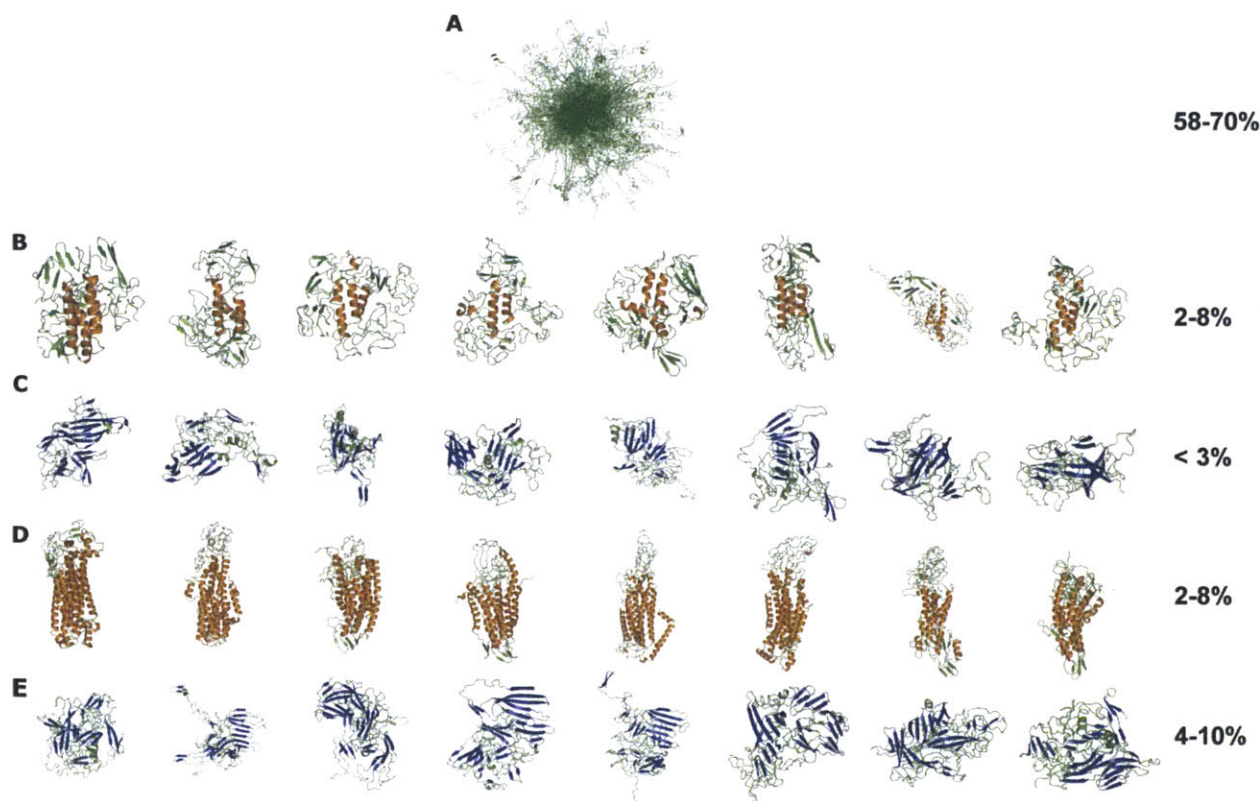


Figure 3. Types of alpha-synuclein structures in our ensemble. Monomers are aligned to each other (A) to demonstrate that they form a structurally heterogeneous set. For the multimeric species, the top 8 structures from each category in terms of secondary structure content are shown: (B) helical-rich trimers; (C) strand-rich trimers; (D) helical-rich tetramers; and (E) strand-rich tetramers.

To determine how each of these multimers may influence alpha-synuclein self-association, we focus on the position and conformation of the subsequence NAC(8-18), which corresponds to the minimal segment of alpha-synuclein that can initiate the formation of toxic beta-strand rich aggregates *in vitro*(154). This is of particular interest because toxic soluble oligomers of alpha-synuclein and other related IDPs contain significant beta-structure(81, 155). 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 $30.6\% \pm 1.0\%$ (Figure 4). For comparison, the solvent exposure of the NAC(8-18) region in the monomeric fraction is $58.6\% \pm 4.2\%$. 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 beta-rich structures, which then progress to form higher order aggregates. In the beta-rich conformers, the NAC(8-18) segment has already been incorporated into beta 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 beta-strand rich multimers.

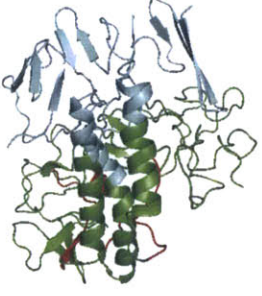

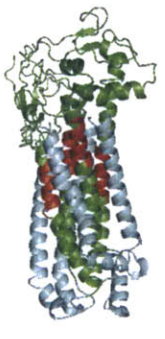

		NAC(8-18) % solvent exposure	N-terminal 1-48 % solvent exposure
A		45.7 ± 1.0%	28.9 ± 0.7%
B		41.1 ± 0.6%	35.6 ± 0.5%
C		30.6 ± 1.0%	34.1 ± 1.0%
D		38.4 ± 1.4%	43.9 ± 1.5%

Figure 4. Normalized solvent accessibility (\pm 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 alpha-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 alpha-synuclein is consistent with this notion(156). More importantly, ^{15}N relaxation experiments performed on monomeric mouse alpha-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 alpha-synuclein – a finding suggesting that secondary structure formation is more prevalent in the mouse form of the protein(157). 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 alpha-synuclein(157). 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 alpha-synuclein in each of these multimeric states, as shown in Figure 3. Helical trimers and tetramers preferentially place the N-terminal region of alpha-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 alpha-synuclein plays a critical role in the formation of helical structures(158-160), 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 4 shows two structures that involve the N-terminal residues in beta-sheet formation, highlighting the beta-strand propensity of these residues.

Interestingly, however, beta-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 alpha-synuclein aggregation *in vitro* (Figure 5). Although it is not clear whether the NAC component or the N-terminal region provides the primary

impetus behind the oligomerization propensity of alpha-synuclein, our data are consistent with a model whereby the initial stages in toxic oligomer formation is the formation of an N-terminal rich beta-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 4, supporting the notion that this structure acts as a non-toxic storage mechanism.

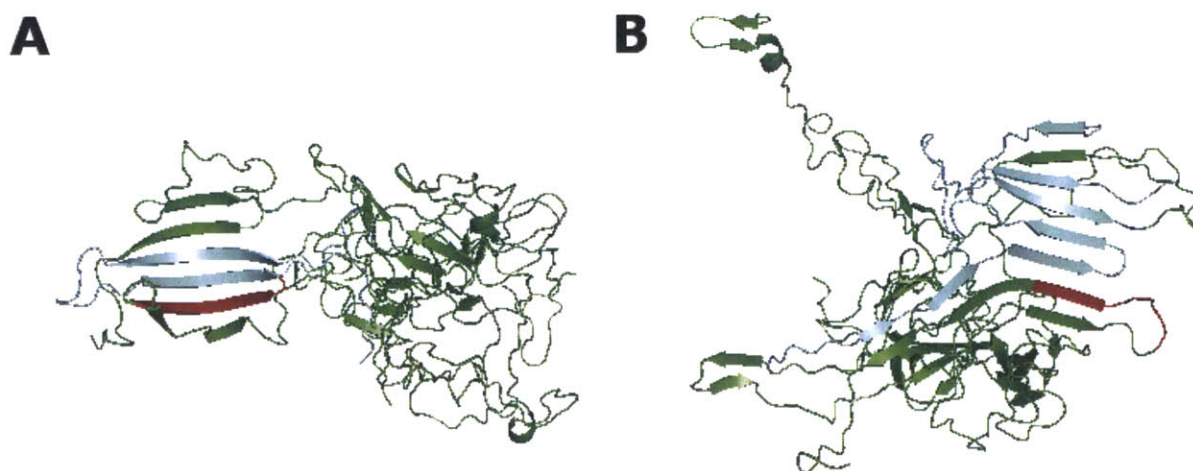


Figure 5. Two representative structures of strand-rich tetramers. The N-terminal residues 1-48 of the monomers participating in sheets are shown in cyan. NAC 8-18 residues participating in sheets are shown in red.

Conclusions

In this study we constructed an ensemble for the multimeric state of alpha-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 alpha-synuclein constructs are isolated from *E. coli*, red blood cells and human neuroblastoma cell lines(92, 134). 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 forms of alpha 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 alpha-synuclein multimers. Given that our previous study suggested that the monomeric alpha-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 alpha-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 6 and 7). 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 (65). These data are then used to calculate ensemble average properties with their corresponding confidence intervals.

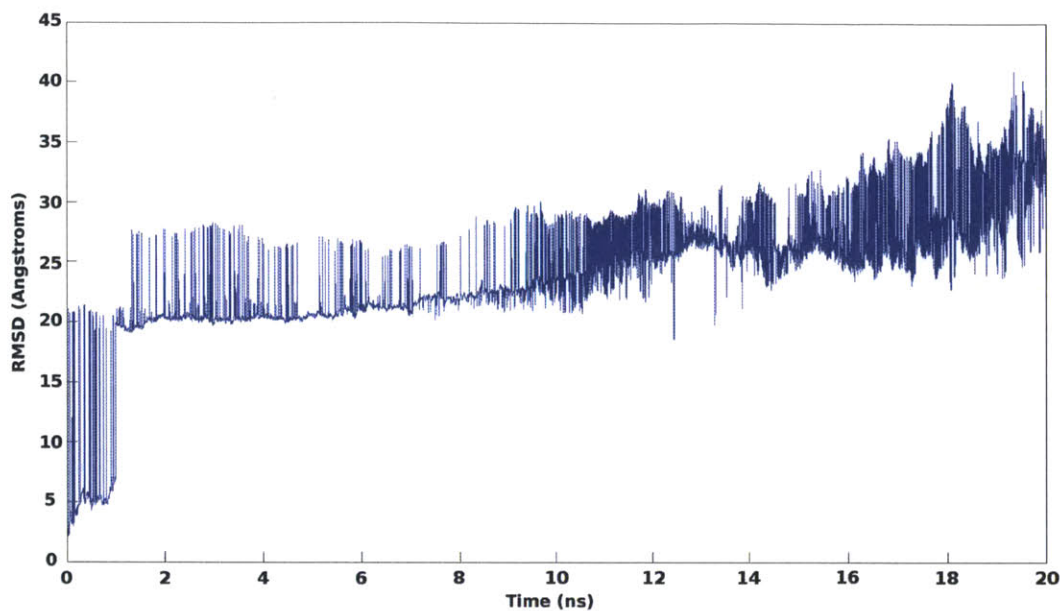


Figure 6. Conformational heterogeneity of a single REMD simulation. Shown is the Ca-RMSD of a structure at time t in the 298K temperature replica compared to the original seed structure (in this case the threaded helical tetramer).

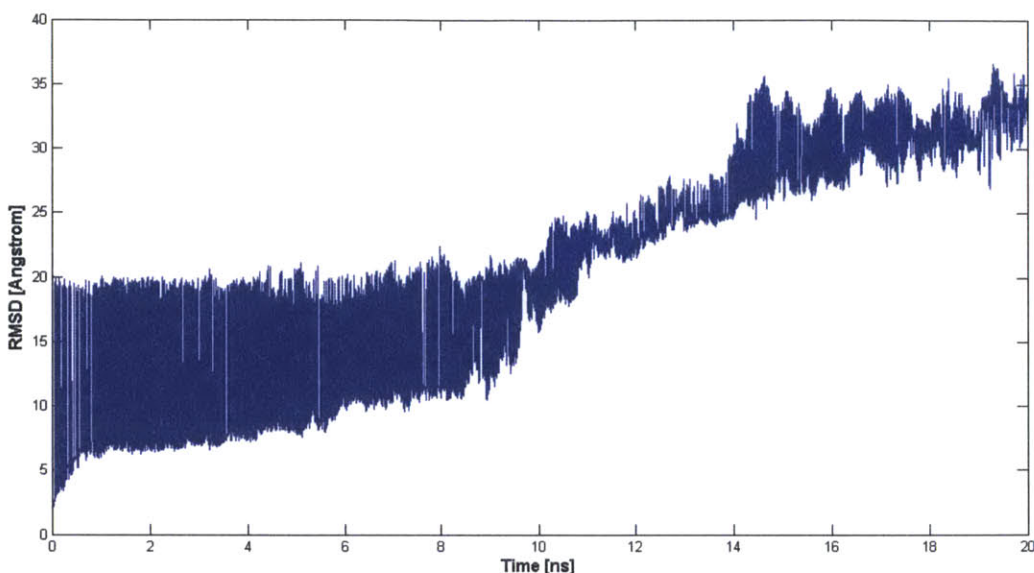


Figure 7. Conformational heterogeneity of a single REMD simulation. Shown is the Ca-RMSD of a structure at time t in the 298K temperature replica compared to the original seed structure (in this case the structure derived from limited NMR data).

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(61, 65). One advantage of the BW formalism is that it has a built in measure of uncertainty, $0 \leq \sigma_{\tilde{w}^B} \leq 1$, that is correlated with model correctness(65). When $\sigma_{\tilde{w}^B} = 0$, we can be relatively certain that the model is correct. By contrast when $\sigma_{\tilde{w}^B} = 1$, it is likely that the ensemble is far from the truth. In the present case, this uncertainty parameter is non-zero: $\sigma_{\tilde{w}^B} = 0.47$. 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 alpha synuclein is predominantly disordered within the crowded intracellular environment(137).

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(92, 133, 161). These latter experiments have led some to conclude that alpha-synuclein predominantly exists as a disordered monomer under physiologic conditions(133). 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(136). 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* alpha-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.

To understand why helical states are aggregation resistant, we focus on the minimal segment, NAC(8-18), needed to initiate alpha-synuclein aggregation *in vitro*(154). 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 beta-strand rich oligomers. In the beta-rich tetramer conformers, the NAC(8-18) segment has already been subsumed in a central beta 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 beta-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 alpha-synuclein suggested that the protein had considerable helical content on average(92, 134). 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(162, 163). For example, we used several different algorithms to quantify the helical content from the published CD spectrum of alpha-synuclein isolated from human red blood cells(92), 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 alpha-synuclein *in vitro* can influence its secondary structure propensity and the precise effect may vary on the post-translational state of the protein(136, 164, 165). Therefore it is not clear whether the published CD spectrum reflects the structure of alpha-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 alpha-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, alpha-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(92). 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 alpha-synuclein(133) and both constructs form tetrameric structures with increased alpha-helical content(92, 134, 136). Lastly, monomeric forms of both constructs have similar aggregation profiles while the tetrameric forms of both constructs do not aggregate(92, 134). These similarities suggest that acetylation of the N-terminal and the 10 residues elongation of the N terminal region in alpha-synuclein serve a similar purpose with regard to their effect on the alpha-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 alpha-synuclein isolated from other cell types, such as neurons or red blood cells, may not be well described by the ensemble presented here.

The Mechanism of Amyloid- β Fibril Elongation

The work presented in this chapter was published in the journal *ACS Biochemistry*, Volume 53 (44), pp 6981-6991, on October 20th, 2014.

Abstract

Amyloid- β is an intrinsically disordered protein that forms fibrils in the brains of patients with Alzheimer's disease. To explore factors that affect the process of fibril growth we computed the free energy associated with disordered Amyloid- β monomers being added to growing amyloid fibrils using extensive molecular dynamics simulations coupled with umbrella sampling. We find that the mechanisms of A β 40 and A β 42 fibril elongation share many features in common, including the formation of an obligate on-pathway β -hairpin intermediate that hydrogen bonds to the fibril core. In addition, our data lead to new hypotheses as to how fibrils may serve as secondary nucleation sites that can catalyze the formation of soluble oligomers – a finding in agreement with recent experimental observations. These data provide a detailed mechanistic description of Amyloid- β fibril elongation and provide a structural link between the disordered free monomer and the growth of amyloid fibrils and soluble oligomers.

Introduction

The amyloid- β ($A\beta$) protein, a 39-42 residue intrinsically disordered peptide(16), has long been implicated in the etiology of Alzheimer's disease.(13) It is formed through directed proteolytic cleavage of the Amyloid Precursor Protein by β - and γ -secretase enzymes(83). $A\beta_{42}$, the 42 residue cleavage product, has been identified as the most prone to forming aggregates, both in the form of low molecular weight soluble oligomers, and insoluble amyloid fibrils.(15, 166-168) While it is most conspicuously deposited in extracellular plaques composed of amyloid fibrils, a growing body of evidence suggests that soluble oligomeric aggregates, rather than fibrillar aggregates, are responsible for the neurotoxicity observed in Alzheimer's disease, leading to a shift in focus away from the latter and towards the former in the effort to combat the disease.(83, 169) Recently, however, it has surfaced that the rate of formation of these oligomeric species of $A\beta_{42}$ is dependent not only on the concentration of available monomeric $A\beta$, but also of amyloid fibrils, suggesting that fibrils act as catalysts for the formation of toxic oligomeric aggregates(76), and reinstating the fibrillar species as a protagonist in the disease process.

The mechanistic details of the interplay between monomers, toxic oligomers and insoluble fibrils remain unknown. Some studies have suggested the existence of a common pathway in which oligomers are on-pathway intermediates in fibril formation, while others propose that oligomers and fibrils are generated through independent pathways.(170) It is likely that therapeutic strategies aimed at stopping the formation of fibrils, which exhibit lower polymorphism than the oligomeric states, will be more tractable in the short- to medium-term.(80) As such, it is of interest to map out the process of transitioning from a disordered monomer of $A\beta$ to a folded amyloid fibril in order to identify key intermediates along the pathway that could form viable therapeutic targets. In this article, we present a detailed computational analysis of $A\beta_{42}$ and $A\beta_{40}$ fibril elongation using atomistic simulations. In-so-doing, we recover several independent experimental observations, integrating them into a common pathway, and

ascribe a critical role to a β -hairpin intermediate in the process of fibril elongation. Furthermore, we find parallels between the process of amyloid fibril elongation in the context of a disordered protein, and the process of globular protein folding in general.

Materials and Methods

Model system

For the A β 42 fibrils, the PDB ID '2BEG' structure was used as a starting point.(58, 110) The structure contains 5 monomers, and was extended to 8 total monomers in the same configuration (by extending the even end of the fibril) to separate the two ends of the fibril by a greater distance. For the A β 40 fibrils, structures for the twofold positive- and negative-stagger fibrils (2LMN and 2LMO in the PDB) were used as starting points.(108) Molecular dynamics simulations were performed using a polar hydrogen model.(144) Atoms within the fibril core are fixed while atoms in the free monomer were allowed to move.

Reaction coordinates and umbrella sampling

The ξ reaction coordinate was computed as the mean of the heavy-atom (N-O) distances between the atoms involved in the intermolecular hydrogen bonds of the A β 42 (110) and A β 40 fibril models (108) (Fig. 1).

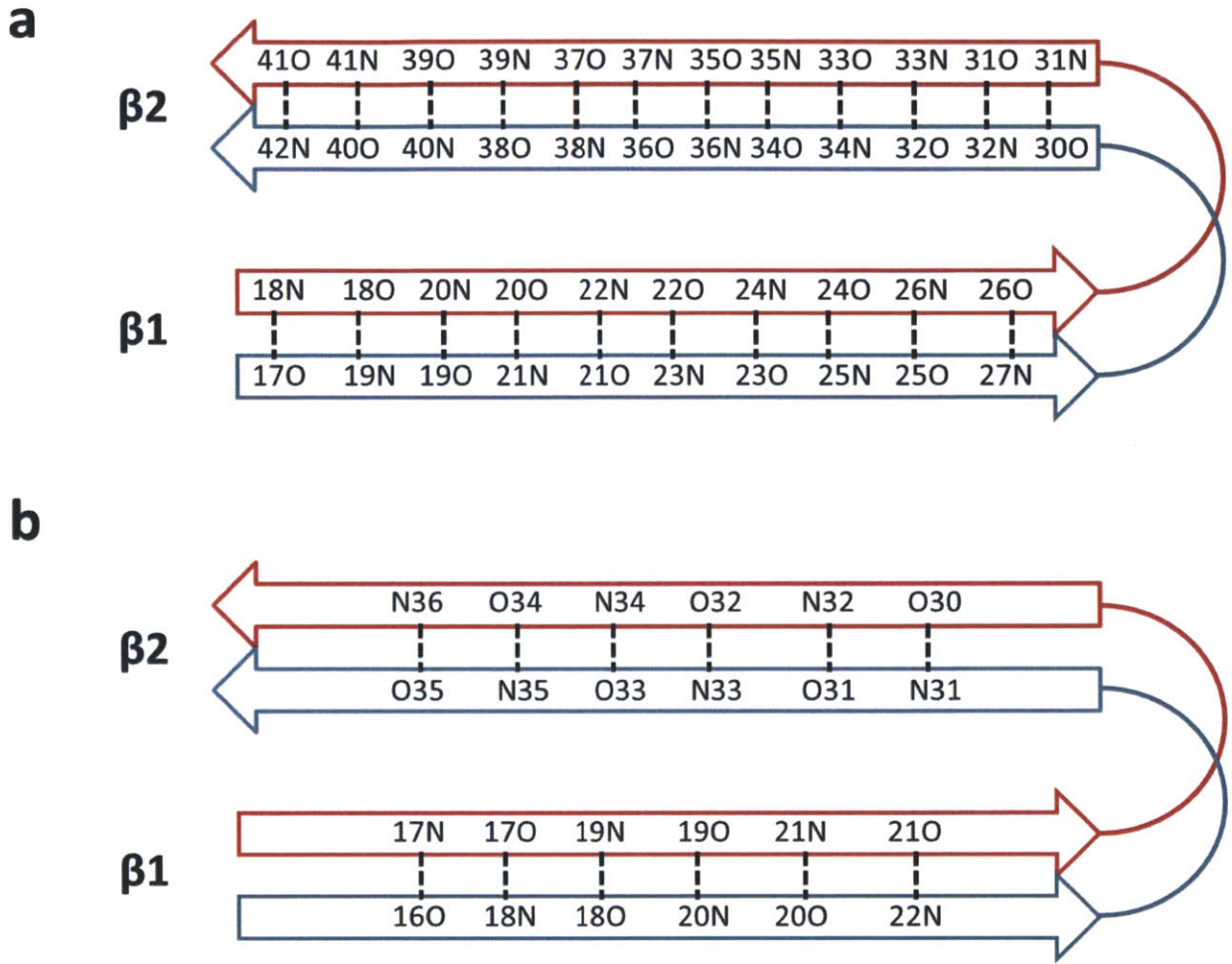


Figure 1. Schematic of the heavy atoms involved in inter-molecular hydrogen bonds in the fibril core. The reaction coordinate ξ is defined as the mean of all the N-O distances depicted. The two strands within a bound monomer are labeled β_1 and β_2 . Heavy atoms defining ξ for the $A\beta_{42}$ fibril are shown in (a), while those for $A\beta_{40}$ fibrils are shown in (b).

We compute f_β according to the continuous function developed by Vitalis *et al.* (12),

$$f_\beta(\phi_1, \psi_1, \phi_2, \psi_2, \dots, \phi_N, \psi_N) = \frac{1}{N} \sum_{i=1}^N f_\beta^{(i)}(\phi_i, \psi_i),$$

where $N=24$ for $A\beta_{42}$ and $N=30$ for $A\beta_{40}$,

corresponding to the number of residues in the system that have both a ϕ and ψ angle (the first and last residues have undefined ϕ and ψ angles, respectively), and

$$f_{\beta}^{(i)}(\phi_i, \psi_i) = \begin{cases} 1 & \text{if } (\phi_i - \phi_{\beta})^2 + (\psi_i - \psi_{\beta})^2 < r_{\beta}^2 \\ \exp(-\tau_{\beta} D_{(i)}^2) & \text{otherwise} \end{cases} \quad (1)$$

where $(\phi_{\beta}, \psi_{\beta})$ is the center of circular region in Ramachandran space with radius r_{β} ;

$D_{(i)}^2 = \left(\sqrt{(\phi_i - \phi_{\beta})^2 + (\psi_i - \psi_{\beta})^2} - r_{\beta} \right)^2$ (the square Euclidean distance between (ϕ_i, ψ_i) and

the boundary of the basin); and τ_{β} is a decay constant. Since differences between

angles can take two possible values due to their periodic nature, any distance in

Ramachandran space, including $D_{(i)}^2$ and angle differences, e.g. $(\phi_i - \phi_{\beta})$, is interpreted

as the minimum possible distance in this space. The center of the basin $(\phi_{\beta}, \psi_{\beta})$ was

defined by $\phi_{\beta} = -152.00^{\circ}$ and $\psi_{\beta} = 142.00^{\circ}$; i.e., the values used by Vitalis et al.(12)

The parameters τ_{β} and r_{β} were chosen to optimize agreement with strand assignments

made by DSSP.(147) More precisely, for a given structure in the PDB one can calculate

f_{β} - which corresponds to the fraction of residues that adopt extended structure

consistent with a beta strand - and one can compute the beta strand percentage using

DSSP. The parameters τ_{β} and r_{β} were chosen to ensure that these calculated values

would be similar for a relatively large set of structures chosen from the PDB. Only

structures consisting of a single chain (with no ligands were used). Moreover, we

ensured that any two structures in the final set had less than 30% sequence identity.

This resulted in 5827 structures. A grid-search was performed for values of τ_{β} between

0.0001deg^{-2} and 0.04deg^{-2} in increments of 0.0001deg^{-2} , and for values of r_{β} between 20°

and 100° in increments of 1° .

In practice agreement with the DSSP strand content values were achieved by minimizing

the objective function $f_{obj} = k_1 \frac{1}{\rho} + k_2 |m-1|$, where ρ is the correlation between the f_{β} and

DSSP scores, m is the gradient of the linear regression of f_{β} against the DSSP-E score,

and k_1 and k_2 are normalizing constants to ensure that the correlation term ($1/\rho$) and

the gradient term ($m-1$) have similar ranges over the dataset. Minimizing this function

yielded final parameters were $\tau_{\beta} = 0.0029\text{deg}^{-2}$ and $r_{\beta} = 62^{\circ}$. With these values we

obtained a correlation of 0.93 between f_β and the DSSP scores, and the corresponding linear regression had a gradient of 0.96.

The final potential energy function is of the form $U_{umbrella} = U_{CHARMM} + U_\xi + U_{f_\beta}$, where $U_\xi = k_\xi (\xi - \xi_0)^2$ and $U_{f_\beta} = k_{f_\beta} (f_\beta - f_0)^2$, (defining a sampling window centered about the values ξ_0 and f_0) and U_{CHARMM} is the CHARMM potential energy function.(171) To use this function for dynamical simulations, the U_{f_β} potential (along with its derivative, which is needed for the force calculations) needs to be added to the CHARMM code. This is described in detail in the Appendix. However, while one can use the form of equation (1) for dynamical simulations, this is not optimal because it is not continuous at the basin boundary, r_β . This leads to unstable trajectories. Consequently, we also developed an alternate form for $f_\beta^{(i)}(\phi_i, \psi_i)$ based on a continuous, two-dimensional Gaussian function for f_β :

$$f_\beta^{(i)}(\phi_i, \psi_i) = \exp\left(-\left(\frac{(\phi_i - \phi_0)^2 + (\psi_i - \psi_0)^2}{2\sigma^2}\right)\right) \quad (2)$$

The center and standard deviation of this alternate form were chosen to again match agreement with calculated DSSP results (as outlined above). Preliminary data suggest that both methods yield similar results, while the latter method had greater numerical stability.

The harmonic force constants k_{f_β} and k_ξ were chosen to obtain adequate overlap between histograms arising from adjacent umbrella sampling windows. Sampling was initiated from reaction coordinate values closest to the fibril model configuration. In the A β 42 model, this corresponds to a value of $f_\beta=0.75$ and $\xi=2.73\text{\AA}$, i.e. $f_\beta=10/13$ and $\xi=3\text{\AA}$. f_β was sampled with values of f_0 between 0 and 1 in increments of 1/13 (~ 0.08) for the A β 42 fibril, which roughly translates into biasing the monomer along its strand contents two residues at a time, while for the A β 40 fibrils, increments of 1/6 (~ 0.17) were used. ξ was sampled between values of $\xi_0=3\text{\AA}$ and $\xi_0=70\text{\AA}$ in both cases. Increments of ξ_0 were spaced by 0.5\AA for $3\text{\AA} \leq \xi_0 < 17\text{\AA}$, and spacing increased to 1\AA

between $17\text{\AA} \leq \xi_0 \leq 42\text{\AA}$, and to 2\AA onwards until $\xi_0=70\text{\AA}$ for A β 42, while for A β 0 increments of ξ_0 were spaced by 0.5\AA for $3\text{\AA} \leq \xi_0 < 10\text{\AA}$ and spacing increased to 1\AA thereafter. k_{fp} was fixed at $350 \text{ kcal mol}^{-1}$. k_{ξ} was set to $1 \text{ kcal mol}^{-1} \text{\AA}^{-2}$ for $3\text{\AA} \leq \xi_0 < 42\text{\AA}$ and to $0.01 \text{ kcal mol}^{-1} \text{\AA}^{-2}$ for $42\text{\AA} \leq \xi_0 \leq 70\text{\AA}$. Initial data suggested that in certain regions, sampling was insufficient in the ξ reaction coordinate. In those cases additional sampling was performed at additional values of ξ (Fig. S7).

The umbrella potential for the ξ reaction coordinate was included by using the RESD (REStrained Distances) command in the CHARMM molecular dynamics package (v36b2).(171) For each pair of values (f_0, ξ_0) (i.e., a “window”) the system was equilibrated for 5ns followed by a 45ns production run (Fig. S8 and Fig. S9). All simulations utilized an implicit model for solvent. While simulations with explicit solvent may provide a more realistic representation of the solvent environment, they result in lengthy production runs to achieve convergence because relaxation of explicit water at each value of the reaction coordinate can be very long. For this reason most umbrella sampling simulations with explicit solvent have been applied to systems that are considerably smaller than that considered in the present study, or have utilized a one-dimensional reaction coordinate.(172-174) To reach the simulation timescales needed to ensure convergence of the reaction coordinates in our umbrella sampling windows, we conducted these simulations using the implicit solvent model EEF1(143) because: 1) prior work suggests that one can obtain free energy profiles (for peptides that form aggregates) with this model that is similar to what would obtain with explicit solvent (175); and 2) other studies which looked at dimerization of peptides that form amyloid precursors suggest that EEF1 produces results that are the closest to experiment (relative to generalized born and analytic continuum electrostatic models).(176)

The initial systems were linearly heated to 310K over 100ps and then coupled to a Nosé-Hoover thermostat of the same temperature.(177) (Bond lengths involving hydrogens were fixed using SHAKE, and simulations were performed using the CHARMM package version 36b2 with a 2fs timestep (171). The total simulation time for both the A β 40 and A β 42 models was $\sim 100\mu\text{s}$. Values of the reaction coordinates from the trajectories were

saved every 5ps, yielding 9,000 data points each per window. The resulting biased probability distributions were recombined to generate the final unbiased PMF using Grossfield's standard implementation of the two-dimensional Weighted Histogram Analysis Method (2D-WHAM). (178, 179)

In addition to the potential of mean force calculations outlined above, we performed unrestrained simulations for both A β 40 and A β 42 models. In each case the setups were identical to the umbrella sampling simulations, except that the systems were linearly heated to 450K, also over 100ps, after which production runs of 1 μ s were performed, and no biasing potentials were employed.

Results

Our calculations began with a model of the A β (17-42) fibril core that was constructed using constraints arising from different experimental observations (e.g., hydrogen/deuterium exchange, mutagenesis studies and solid-state NMR) (PDB ID 2BEG).(58, 110) This structure is composed of two intermolecular, in-register and β -sheets formed by residues 18-26 (strand β 1) and 31-42 (strand β 2), with the first 16 N-terminal residues being disordered and external to the fibril.(110) To study the process of fibril elongation, we began with this model that we refer to as the "fibril core" (consisting of residues 17-42), and calculated the free energy for the folding of an A β monomer being added to the odd end of the fibril core (Fig. 2A). In practice, the free energy calculations begin with the folded state (i.e., with the A β 42 monomer already bound to the fibril core) and the free energy profile for unbinding (or unfolding) of an A β 42 monomer is calculated, thereby enabling the simulations to begin with a well-defined structure. Since the free energy itself is a state function, the final free energy surface, in principle, is not determined by the order in which the calculations proceed.

The free energy associated with folding was calculated as a function of two reaction coordinates: the average heavy-atom (N-O) distance between pairs of atoms that form inter-molecular hydrogen bonds between the A β monomer and the strands at the odd end of the fibril core, ξ , and the fraction of residues in the A β monomer adopting β -strand secondary structure, f_β . The former acts as a proxy for measuring the distance of

the A β monomer from the fibril (Fig. 2B), and the latter ensures that we sample a wide range of β -strand content as the monomeric disordered protein folds to the fibrillar state (Fig. 2C) (91). The resulting free energy surface (FES) is a function of these two reaction coordinates, and is also referred to as a potential of mean force.(180)

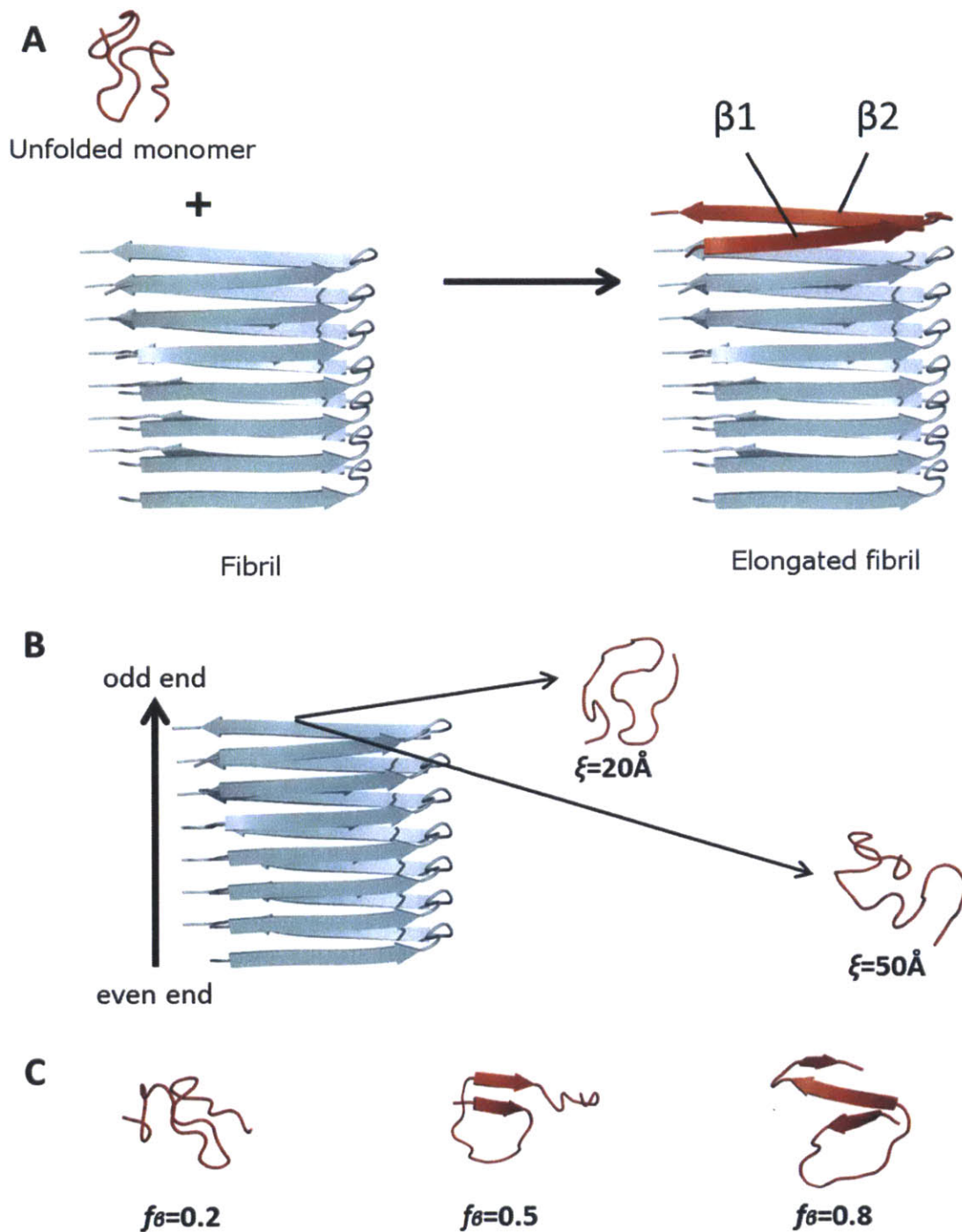


Figure 2. Schematic of the elongation reaction and the reaction coordinates used. The backbone trace of the $A\beta$ monomer is shown red while the backbone trace of the “fibril core” is shown in cyan. Strands in the bound state (A) are labeled as $\beta 1$ (N-terminal) and $\beta 2$ (C-terminal). Different values of ξ and f_β are shown to illustrate the reaction coordinate and the even and odd ends of the fibril core (B and C respectively).

To construct a free energy surface we use umbrella sampling with implicit solvent. The approach requires us to generate continuous “umbrella” potentials to ensure that a wide range of values for the reaction coordinates are sampled.(180) A simple harmonic function is used to restrain the average N-O distance to any desired value. To construct a continuous function that would restrain the fraction of residues that would be in a β -strand conformation, f_β , we used a previously developed formalism that defines a β -basin in dihedral angle space(12) (cf. Methods). Once this function is specified, a simple harmonic function can also be applied to restrain the system to any particular β -strand content. By restraining the simulation to specified values of ξ and f_β we can ensure that a wide range of conformational space is sampled. Moreover, while sampling is employed on a biased energy surface, the final calculated free energy is independent of the choice of the umbrella potentials chosen and therefore represents the true free energy profile calculated from an unbiased energy surface.(179, 180)

The resulting free energy surface has a broad global energy minimum corresponding to the bound, fibrillar state, F (Fig. 3). From these data we identify the minimum free energy path from the unbound state to the bound, folded state (centered about $\xi=3\text{\AA}$); i.e., the path that minimizes the work associated with moving from one state to the other (Fig. 3). We can also identify paths that are within $3kT$ of the lowest energy path, thereby providing information about the diversity of states that are sampled as the system proceeds from the unfolded to the folded state.

Low energy paths connecting the unfolded state to the final folded state have many features in common. To illustrate this we can identify 6 states that are sampled along

the lowest energy path and corresponding structures that are at least 3kT from these states (a, b, c, H, T* and F (Fig. 3). As the monomer approaches the fibril core it preferentially makes contacts with residues 25-31 in the fibril that form turns between the strands (Fig. 4, state a). Essentially, the monomer “rolls” along the fibril making non-specific interactions between the monomer and exposed side-chain of Asn 27, as well as the backbone carbonyl of Lys 28 and backbone nitrogen of Ala 30 (Fig. 5). Association with the odd end of the fibril then ensues, followed by the formation of a strand (β 1) at the odd end (Fig. 4, state b) with the release of roughly 15kcal/mol (Fig. 3). While the β 1 strand (residues 18-26, Fig. 1) of the monomer remains bound at the odd end, the C-terminal residues sample a range of distinct conformations, some of which continue to make contact with the turn residues in the fibril core. Progressive formation of intramolecular hydrogen bonds eventually leads to the formation of structures that contain a β -hairpin (Fig. 4, states c and H). Interestingly, structure (c) contains turns at residues found to be solvent exposed in amide exchange NMR experiments performed on A β 42 oligomers, and closely resembles the proposed structure in that same study, highlighting the possibility that soluble oligomers and fibrils share common intermediates to their formation.(98) To reach the final folded state from H, a free energy barrier of approximately 4kcal/mol has to be overcome. In the associated transition state T*, intramolecular hydrogen bonds between strands β 1 and β 2 in the hairpin are broken (Fig. 4 and Fig. 7). In the final state, these intramolecular interactions are replaced with intermolecular hydrogen bonds between the A β monomer and the fibril core (Fig. 7).

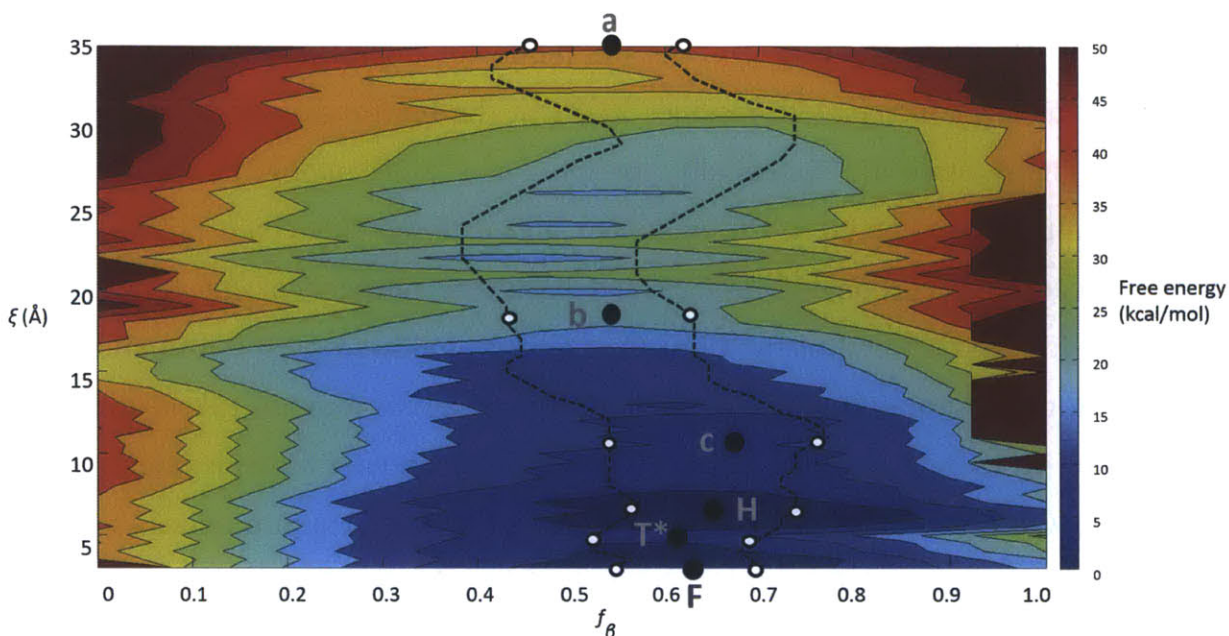


Figure 3. Contour plot of the free energy surface of monomeric A β (17-42) as a function of proximity to the fibril bound state (ξ) and β strand content (f_β), shown for values of ξ smaller than 35Å. Points (a, b, c, H, T and F) along the minimum energy path between the unbound and the bound state are explicitly shown. Dotted black lines represent a $3kT$ envelope around the minimum energy path. Only free energies of states that have an average N-O distance between the A β monomer and the fibril core less than 35Å are shown. The full PMF can be found in Figure 6.*

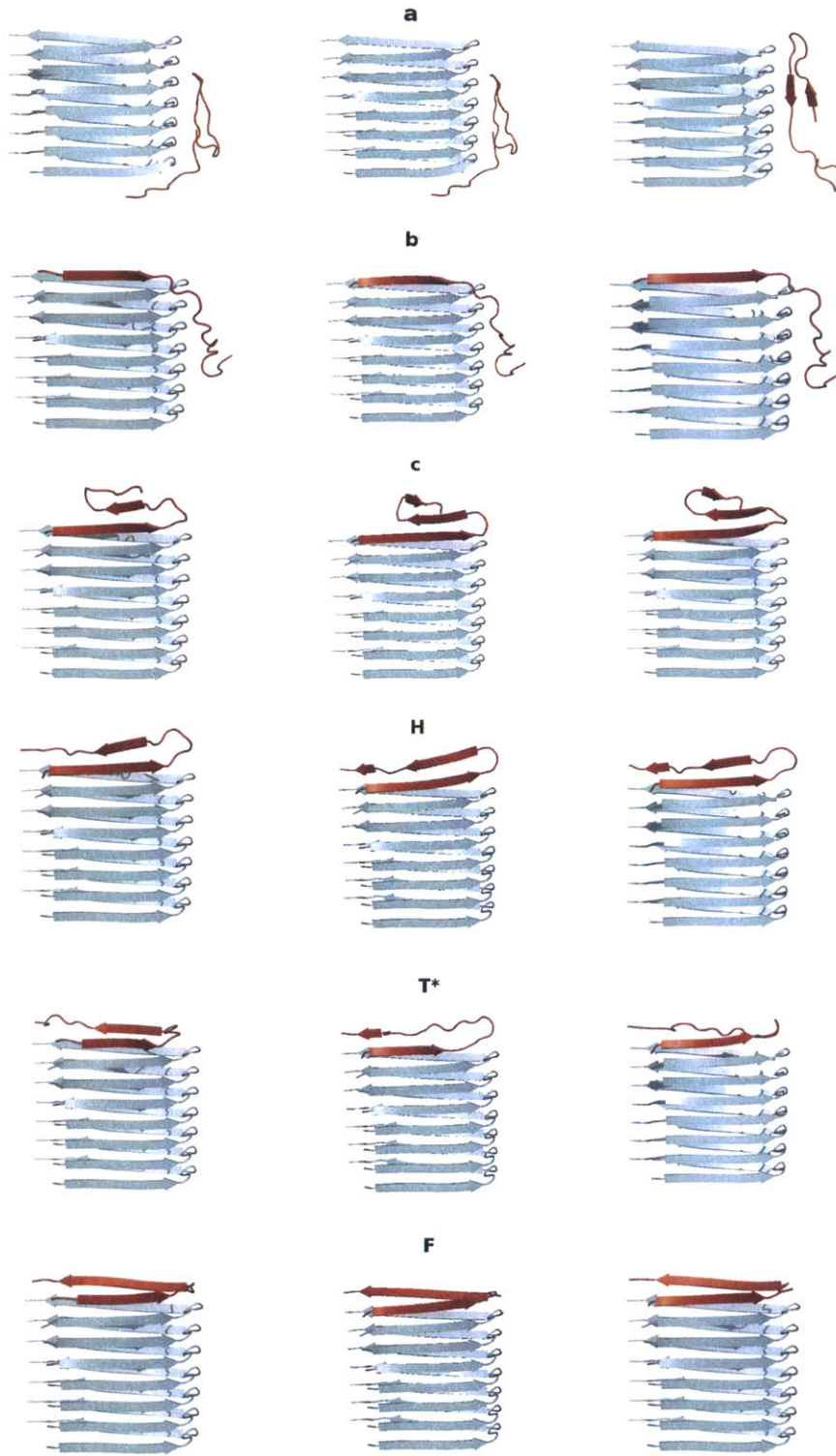


Figure 4. Structures for states a, b, c, H, T* and F. The middle structures are on the lowest energy path and flanking structures have free energies that are at least $3kT$ higher.

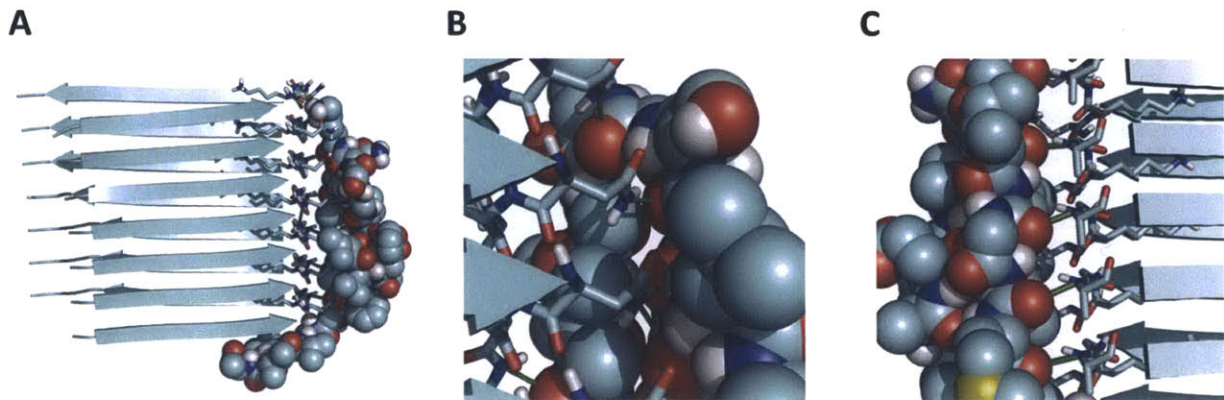


Figure 5. Example of non-specific interactions between the unbound monomer (shown in Van der Waals representation) and the turn between strands $\beta 1$ and $\beta 2$ in the fibril core (shown in liquorice representation) for $33\text{\AA} < \xi < 42\text{\AA}$. The side-chain of Asn 27, as well as the backbone carbonyl of Lys 28 and backbone nitrogen of Ala 30, are exposed to solvent, and are therefore available for forming hydrogen bonds with an unbound monomer. Hydrogen bonds involving the Asn 27 side-chain are shown in panel B, while those involving the backbone nitrogen of Ala 30 are shown in panel C. All hydrogen bonds shown in green have heavy atom distances smaller than 3.5\AA .

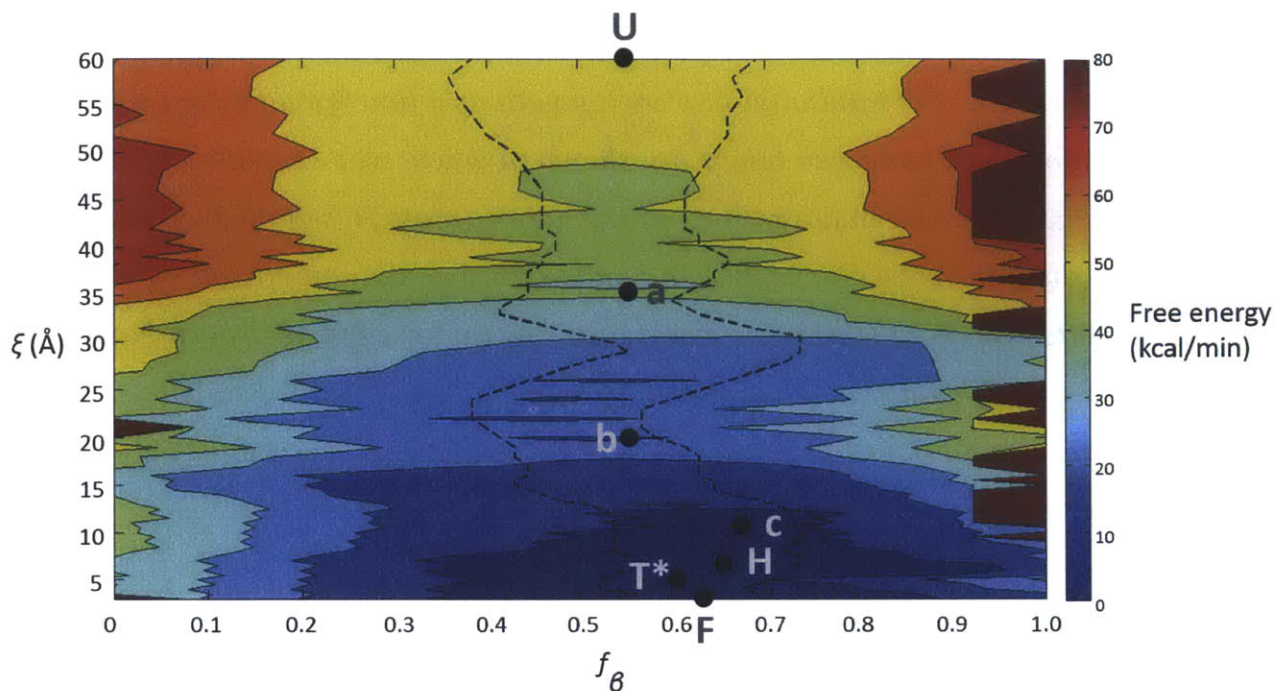


Figure 6. Contour plot of the PMF for the folding of A β (17-42) as a function of proximity to the fibril-bound state (ξ) and β -strand content (f_β). Points (U, a, b, c, H, T* and F) along the minimum energy path between the unbound and the bound state, which are mentioned at several points in the text, are explicitly shown. Dotted black lines represent a $3kT$ envelope around the minimum energy path.

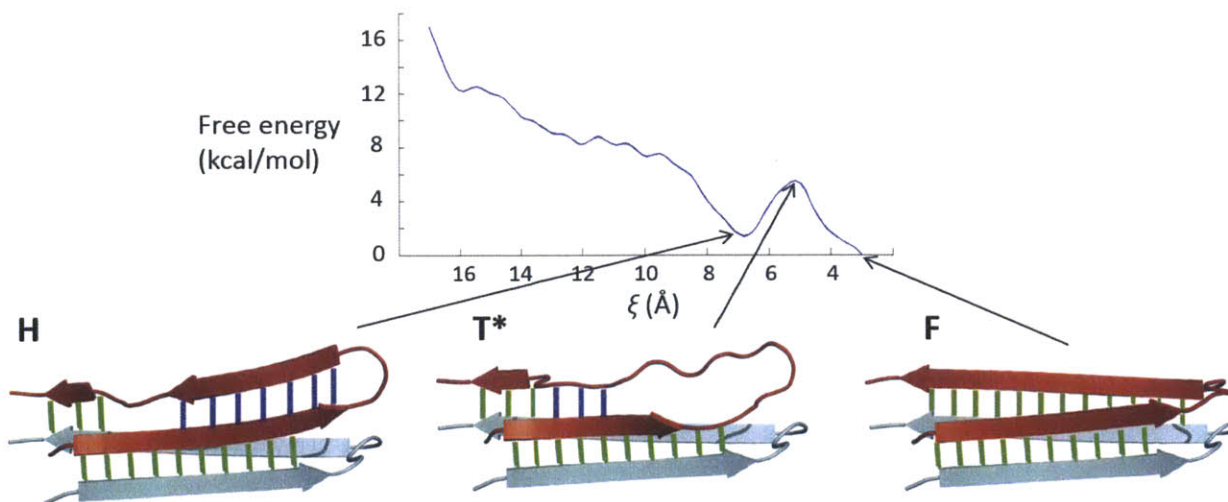


Figure 7. Transition from hairpin structure (H) through the transition state (T) to the fibrillar state (F). The minimum free energy path as a function of ξ for $\xi < 17\text{\AA}$ is shown. Intermolecular hydrogen bonds are shown in green, while intramolecular hydrogen bonds are shown in dark blue. In the hairpin state H, strand β_1 (residues 18-26) forms the intermolecular hydrogen bonds with the adjacent strand in the fibril core, while β_2 strand (residues 31-41) forms intramolecular hydrogen bonds with β_1 . In the transition state T*, most of the intramolecular hydrogen bonds between strands β_1 and β_2 are broken, and intermolecular hydrogen bonds between β_2 and the fibril core form. In the bound state F, all hydrogen bonds are intermolecular.*

To explore how our findings depend on the choice of starting structure, and A β sequence, we recomputed free energy surfaces using a different structure for the fibril core. Since a number of fibril structures in the structural database are composed of two or more filaments (108, 109, 122), we chose two structural models of A β 40 fibers that were built using experimental constraints arising from solid-state NMR experiments. While both structures consist of residues 9-40 as the N-terminal 8 residues were disordered and contain two filaments, they differ with respect to the relative positions and orientations of the β -sheets (PDB ID 2LMN and 2LMO, respectively). (58, 108) In particular, the restraints used to construct these fibril structures were compatible with both positive- and negative-stagger models and therefore two models could be built from the data. (108) However, recent computational studies on models of A β 40 fibrils suggest that only the negative stagger can form left-handed helical superstructures – the twist that has been observed in scanning electron microscopy studies of amyloid superstructures. (112, 113) Indeed our own data are consistent with these observations as we find that the global free energy minimum for the A β 40 structure with positive stagger is not the fibrillar state (Figs. 8 and 9).

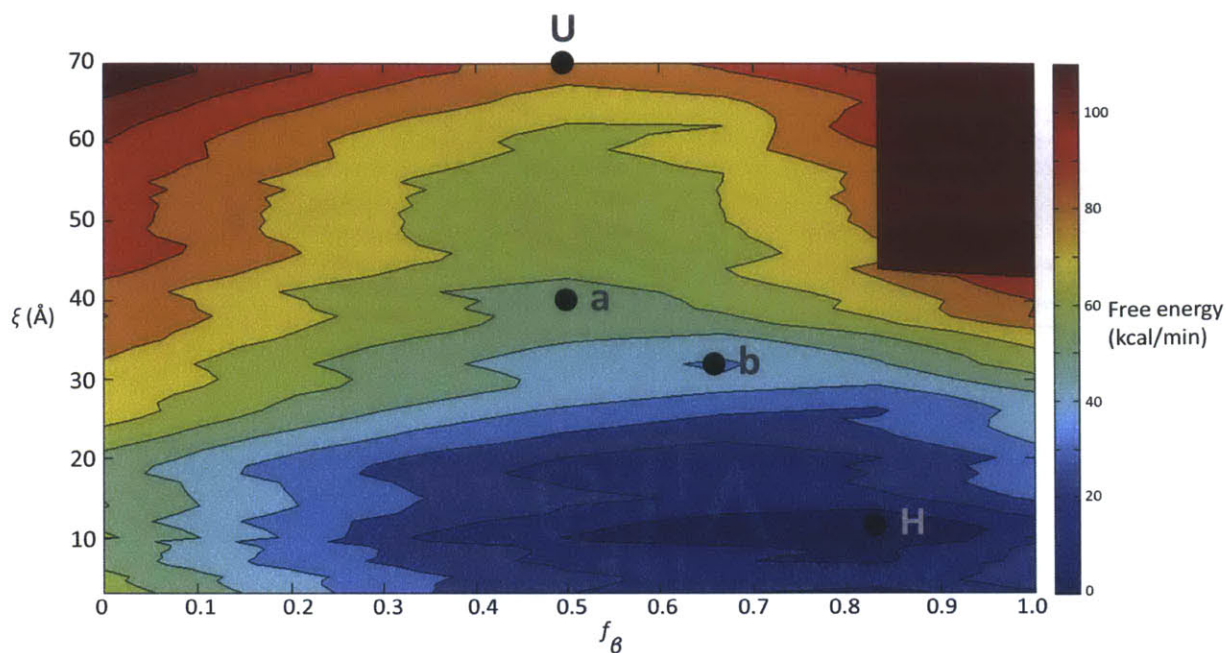


Figure 8. Contour plot of the PMF for the association of $A\beta(9-40)$ with a twofold-symmetric, positive-stagger fibril as a function of proximity to the fibril-bound state (ξ) and β -strand content (f_β). Points (U, a, b and H) along the minimum energy path between the unbound and the hairpin state are explicitly shown. Note that the global free energy minimum corresponds to the hairpin state.

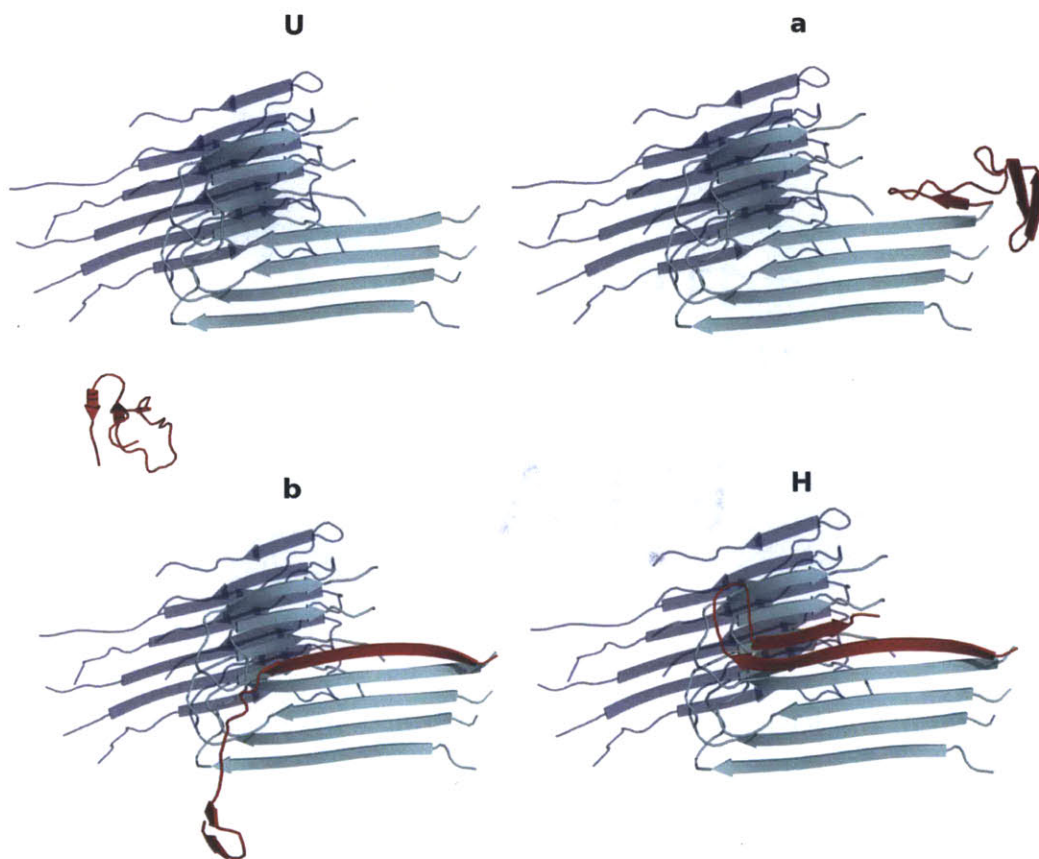


Figure 9. Representative structures from the folding pathway of A β (9-40) on a positive-stagger twofold symmetric fibril. Structures were taken from the umbrella sampling windows corresponding to free-energy minima along f_{β} at values of ξ labeled in Fig. S5. The lowest potential energy structure within that umbrella sampling window was taken as the representative.

The free energy surface for the negative-stagger, twofold-symmetric A β 40, fibril has a broad global energy minimum corresponding to the bound, fibrillar state (Fig. 10). As before, we identified the minimum free energy path from the unbound state to the bound, folded, state (centered about $\xi=3.5\text{\AA}$) (Fig. 10), along with paths that are at least $3kT$ from the minimum energy path.

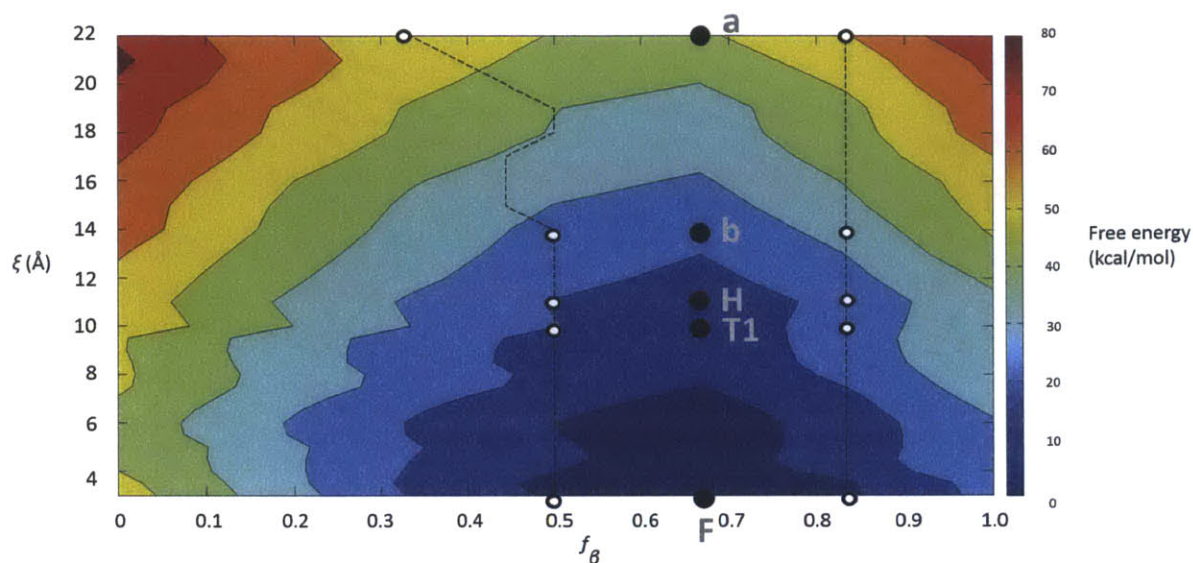


Figure 10. Contour plot of the PMF for the folding of $A\beta(9-40)$ as a function of proximity to the fibril bound state (ξ) and β strand content (f_β), shown for values of ξ smaller than 22\AA . Points (a, b, H, T1 and F) along the minimum energy path between the unbound and the bound state are explicitly shown. Dotted black lines represent states that are at least $3kT$ from the lowest energy path. Only free energies of states that have an average N-O distance between the $A\beta$ monomer and the fibril core less than 22\AA are shown. The full PMF can be found in the Supporting Information.

Low energy paths from the unbound state to the bound, folded, state share many features in common with one another. The incoming monomer initially interacts with the second filament of the fibril (state a, Figs. 10 and 11). The N-terminal $\beta 1$ strand then associates with the odd end of the first filament (state b, Figs. 10 and 11). Next, the $\beta 2$ strand forms intramolecular hydrogen bonds with the $\beta 1$ strand, forming a hairpin intermediate, H (Fig. 11, state H). The reaction then proceeds through several states in which the intramolecular $\beta 1$ - $\beta 2$ hydrogen bonds break and are replaced with intermolecular hydrogen bonds between adjacent $\beta 2$ strands at the odd end of the fibril, ending in the fibrillar conformation, F (Fig. 12).

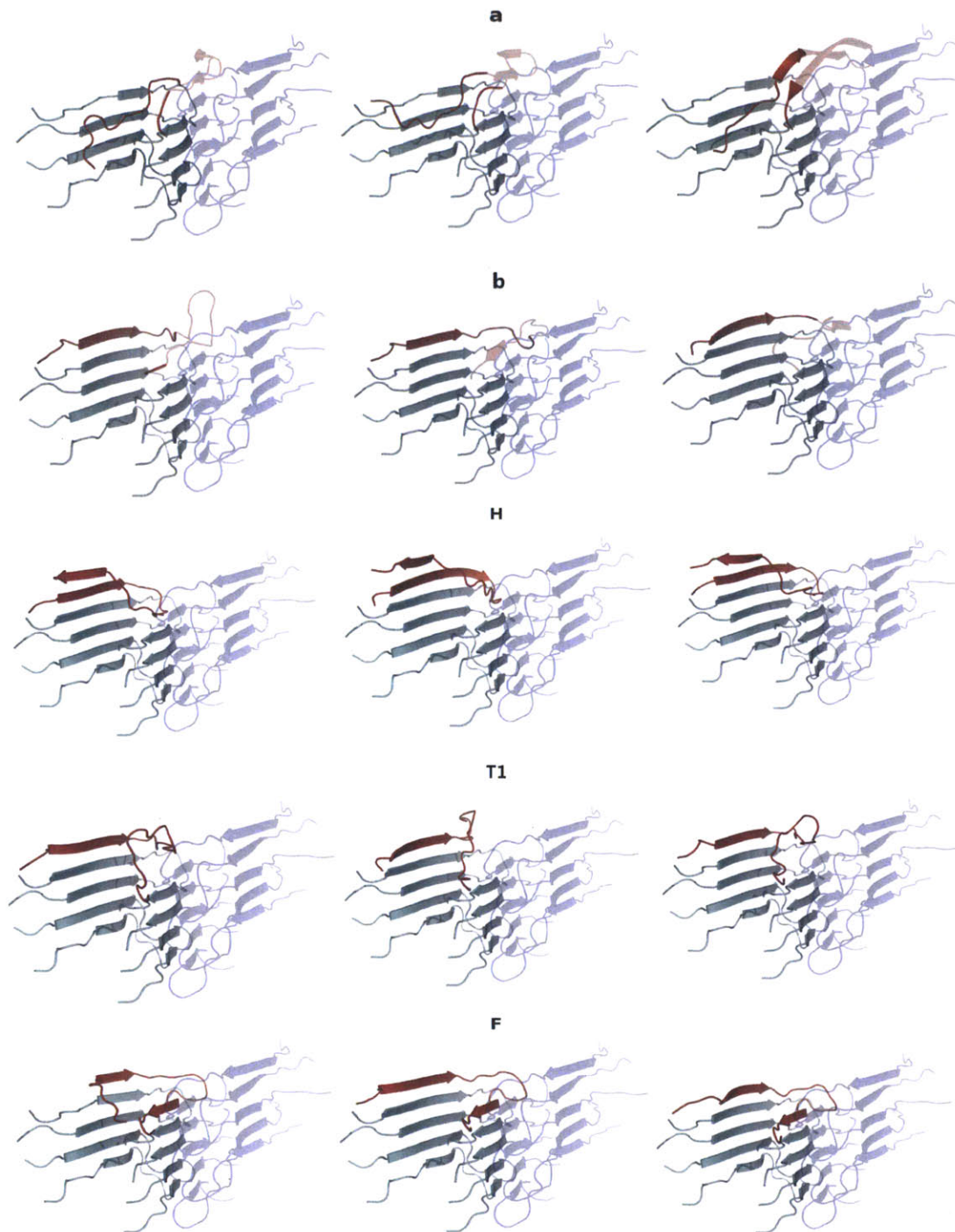


Figure 11. Structures for intermediate states *a*, *b*, *H*, *T1* and *F*. The AB40 structure is composed of two filaments. The first filament is shown in cyan and the second is shown in transparent blue. Regions of the incoming monomer (red) that are associated with the second filament are shown as transparent as well. The middle

structures are on the lowest energy path and flanking higher energy structures having free energies that are at least 3kcal/mol higher.

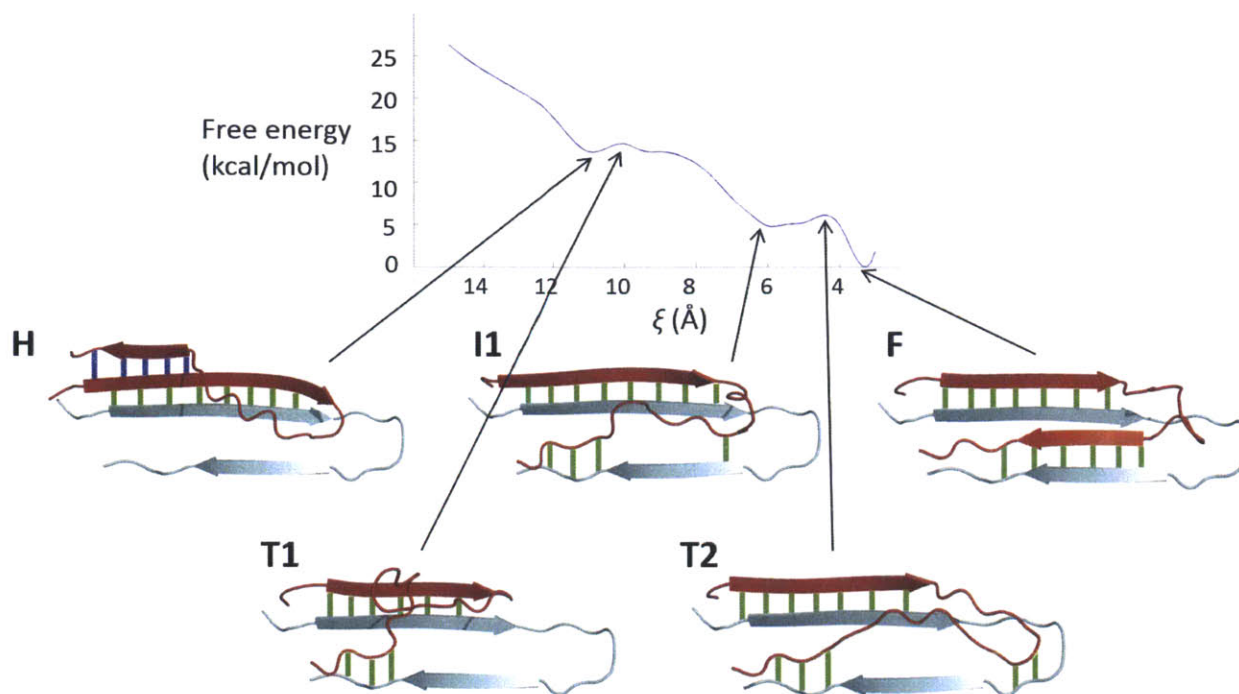


Figure 12. Transition from hairpin structure (H) through the transition states (T1 and T2) to the fibrillar state (F). The minimum free energy path as a function of ξ for $\xi < 15\text{\AA}$ is shown. Intermolecular hydrogen bonds are shown in green, while intramolecular hydrogen bonds are shown in dark blue.

Calculating a free energy surface requires one to pre-specify a set of reaction coordinates. While the calculated free energy difference between the unbound and bound (folded) state is independent of the path, the intermediates sampled along the path will depend on the choice of the reaction coordinates. To test whether the observed intermediates are artifacts of the choice of reaction coordinates, we performed unbiased unfolding simulations of both the A β (17-42) and the A β (9-40) fibril core models. Although these simulations do not allow us to calculate precise free energy differences between states

(as opposed to the detailed free energy simulations described above), they do allow us to probe the dynamics of unrestrained monomers as they unbind, without any bias introduced by a pre-specified choice of reaction coordinate.

We performed a 1 μ s unbiased simulation at 450K on the same A β (17-42) fibril core used in the umbrella sampling simulations, again with the EEF1 implicit solvent model. Over the course of the unfolding simulation, we observe a transition from the fibrillar state, F, to a hairpin state, H, in which strands β 1 and β 2 form intra-molecular hydrogen bonds, which are then broken upon returning to the fibrillar state (Fig. 13). This is consistent with an intermediate, hairpin state H which has a lower stability than F but which is within thermal reach of F – observations in agreement with the lowest energy path on our free energy surface (Fig. 7). Similar unbiased simulations at 450K of the A β (9-40) fibril core were performed, again using the EEF1 implicit solvent model. Conformations sampled during the simulation are consistent with those derived from the free energy surfaces (Fig. 14). More specifically, the trajectory proceeds from the fibrillar state F by breaking the inter-molecular hydrogen bonds of the β 2 strand via states T1 and I1, followed by association of the β 2 strand with the second filament of the fibril structure (state b) and subsequent dissociation – data in good agreement with states observed on the calculated free energy surface (Fig. 12). We note that no hairpin state was significantly sampled during the course of our A β 40 simulations – a finding consistent with the observation that this state has a relatively high energy on the free energy surface.

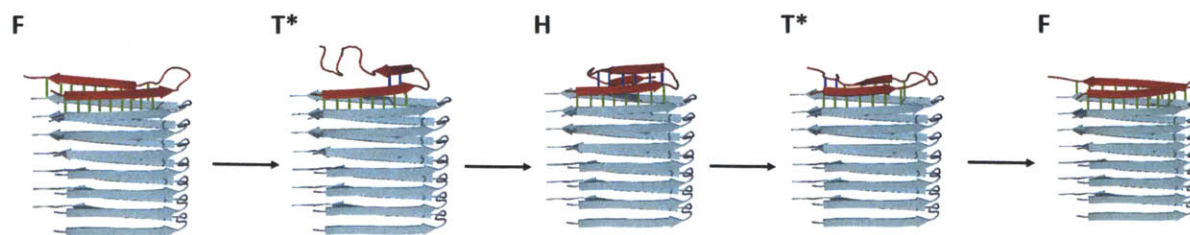


Figure 13. Structures sampled during the 1 μ s unfolding simulation at 450K of the A β (17-42) fibril core. The states are presented in chronological order, and were sampled at the following times: 500ns, 550ns, 600ns, 700ns and 850ns.

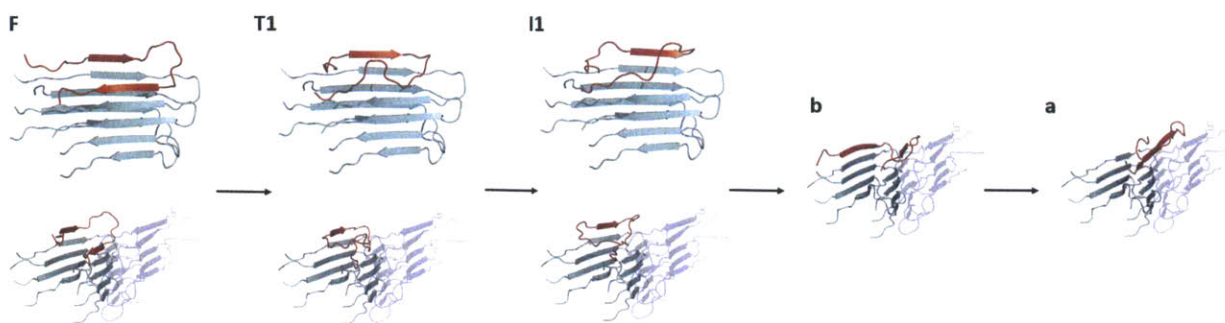


Figure 14. Structures sampled during the 1 μ s unfolding simulation at 450K of the A β (9-40) fibril core. The states are presented in chronological order, and were sampled at the following times: 0ns, 50ns, 125ns, 350ns, and 550ns.

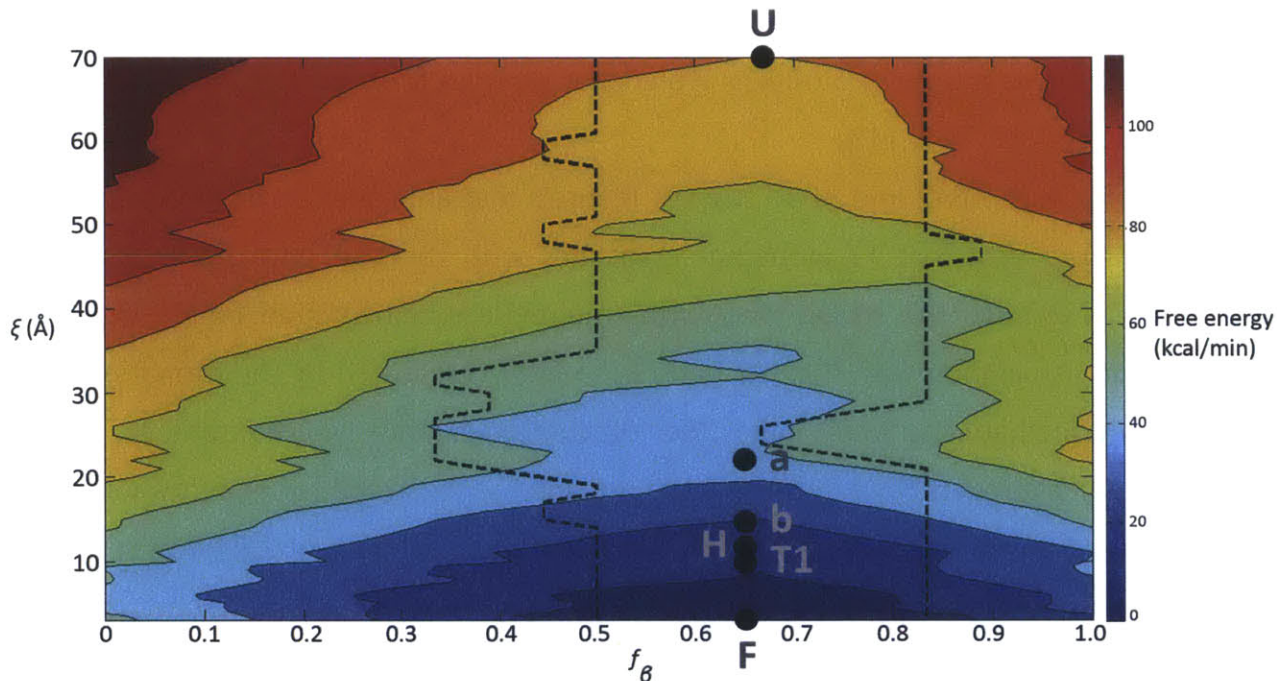


Figure 15. Contour plot of the PMF for the association of A β (9-40) with a twofold-symmetric, negative-stagger fibril as a function of proximity to the fibril-bound state (ξ) and β -strand content (f_β). Points (U, a, b, c, H, T1 and F) along the minimum energy path between the unbound and the bound state are explicitly shown. Dotted black lines represent a 20 kT envelope around the minimum energy path. Note that the global free energy minimum corresponds to the fibrillar state F.

Discussion

A β aggregation lies at the heart of Alzheimer's disease pathology, in the form of amyloid fibrils and lower molecular weight soluble oligomers. Through extensive umbrella sampling simulations performed on experimentally-derived models of fibril structures, we compute a free energy surface for the process of A β 42 and A β 40 fibril elongation. We find that fibril elongation occurs on downhill free energy pathways, ending in the fibrillar conformation, F, which corresponds to the global minimum on the free energy surfaces of both A β 42 and negative-stagger A β 40 fibrils, but not in positive-stagger A β 40 fibrils. The inability of the positive-stagger A β 40 fibril to elongate is consistent with prior data that suggest that positive stagger filaments cannot adopt the superstructural helical twist that has been observed in scanning electron microscopic studies of amyloid fibrils; i.e., positive-stagger protofibrils would not grow to form mature fibrils with the correct helical twist. (112, 113)

Our results for both A β 42 and A β 40 suggest features that are common to the elongation process for both proteins: 1) monomer associates with the odd end of the fibril by forming an N-terminal β 1 strand that forms intermolecular hydrogen bonds with the fibril core; 2) association with the odd end of the fibril is followed by the formation of a common intermediate, H, which takes the form of a β -hairpin where strand β 1 forms intermolecular hydrogen bonds to the fibril core and the β 2 strand forms intramolecular hydrogen bonds with the β 1 strand; 3) disruption of the intramolecular hydrogen bonds within the hairpin leads to formation of the final bound state where the monomer only forms intermolecular hydrogen bonds with the fibril core. For both sequences, a β -hairpin is an obligate intermediate on the folding pathway. These data are consistent with the observation that sequestration of a β -hairpin conformation of A β 40 slows aggregation.(97) Additionally, stabilizing the bend between the two beta strands leads to a significant increase in the rate of fibrillogenesis – a finding also consistent with our results.(181)

Several studies suggest that aggregation-prone states are sparsely populated in the absence of fibril cores, and that stabilization of these states leads to an increase in the

rate of fibril formation.(181-183) Indeed, in a previous study we generated structural ensembles for A β 42 and A β 40, in the absence of a fibril core, using a number of experimental observables as a guide, and observed that β -hairpin conformations were infrequently sampled for both A β 42 and A β 40.(121) Although these states are not highly populated, an analysis of the ensembles suggests that A β 42 is approximately 10 times more likely to adopt β -hairpin structures relative to A β 40. Taken together, these data help to explain why A β 42 forms fibrils much faster than A β 40; i.e., A β 42 is more likely to populate intermediates along the folding pathway. This observation becomes even more pronounced in the presence of the fibril core. When the fibril is present, β -hairpin structures for A β 42 have energies that are only a few kT higher than that of the native, folded, state while β -hairpin structures A β 40 have energies that are significantly higher than the native state energy (Figs. 7 and 12). These data highlight at least one mechanism whereby the presence of fibrils can accelerate fibrillogenesis. More precisely, when fibrils are present, some A β isoforms may be more prone to adopt aggregation-prone structures that can be incorporated into a growing fibril.

A number of studies have attempted to isolate key molecular features involved in fibril or oligomer growth of A β 40 or smaller amyloidogenic peptides derived from the A β 40 sequence.(184-186). A common feature that arises from these studies is that addition of monomer to β -rich template representing either a soluble oligomer or a protofibril, occurs via a “dock-lock” mechanism that is similar to the scheme originally proposed by Esler et al. (187). Docking consists of an incoming monomer loosely associating to the template in a manner such that it can readily dissociate. Locking involves the formation of hydrogen bonds to the template, yielding a structure where monomer dissociation is unlikely. In our studies, monomer initially interacts with the template via non-specific interactions (Fig. 4 and 11, state a) that can involve regions other than the odd end of the fibril. Locking (a relatively slow process) occurs when the β 1 strand of the incoming monomer binds to the odd end of the fibril. Subsequent structural rearrangements in the monomer lead to the formation of the final folded structure.

The free energy surfaces were calculated as a function of two reaction coordinates: ξ , the average N-O distance between the free monomer and the odd end of the fibril; and f_β , the fraction of residues that have phi-psi angles that are consistent with a beta strand. Since the addition of a monomer to the odd end of the fibril can be viewed as a ligand binding reaction, where the ligand changes its structure upon binding, we chose a reaction coordinate, ξ , which quantifies the distance of the monomer to the fibril core, and that ensures that the monomer samples states that have the correct hydrogen bonding pattern. The second reaction coordinate, f_β , which quantifies the β -strand-content of the incoming monomer, ensures that we sample a variety of β -strand-content at any given distance from the fibril core. Indeed, similar reaction coordinates have been used to study ligand binding and protein association.(175, 188-190) Nevertheless, it is important to note that while the calculated free energy difference between the unbound and bound (or folded) state is a function of state, and therefore independent of the path chosen to go from the unbound to the bound state, the observed intermediates are dependent on the choice of reaction coordinates. Although the relatively lengthy simulation time for these calculations ($\sim 100 \mu s$ total for both the A β 40 and A β 42 models) enables the system to sample a variety of different structures during the umbrella sampling runs (e.g., Fig. 16a and b), the choice of the reaction coordinates will influence the structures sampled on the lowest energy path.

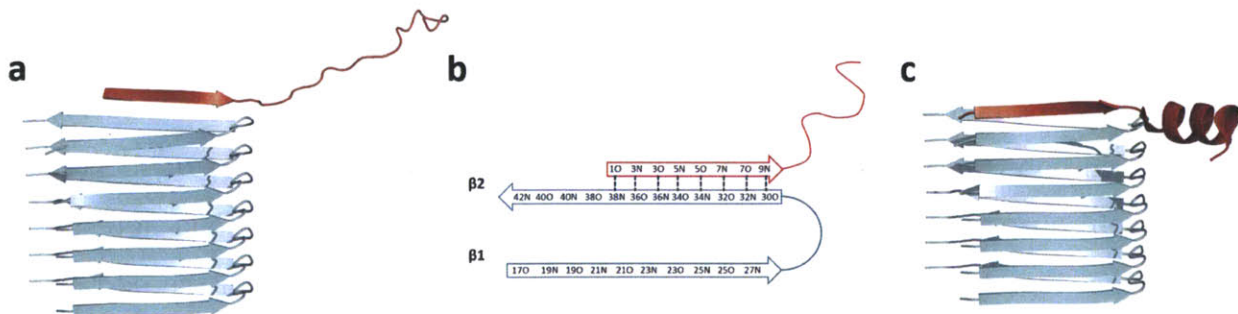


Figure 16. Examples of structures sampled outside the minimum energy path. While the ξ reaction coordinate is a function of distances between atoms that are in close contact in the hydrogen bonds register adopted in the fibrillar state F , alternate

registers can be sampled even at small values of ξ_0 , as shown in the negative stagger A β (9-40) fibril structure in (a), taken from the window centered about $\xi_0=3.0\text{\AA}$ and $f_0=1/6$; (b) Hydrogen bonding pattern associated with figure shown in part (a); (c) While the f_β reaction coordinate we employed only explicitly considers β -strand-content, helices were nonetheless sampled in windows centered about a low f_0 , as in the A β (17-42) (structure taken from the window centered about $\xi_0=22\text{\AA}$ and $f_0=4/13$).

Ideally one could gain insight into the folding pathway from unbiased simulations starting from the unfolded (unbound) state under conditions where the folded state is stable. Since folding occurs on time scales that are typically beyond the reach of atomistic simulations, such simulations are typically not tractable. Information about the folding pathway can sometimes be garnered from unfolding simulations, where folded protein structures are subjected to conditions where the unfolded state is the most stable; e.g., simulations at high temperatures. While it is clear that unfolding at high temperatures is different than folding (which occurs at lower temperatures), a number of studies suggest that high temperature unfolding simulations can capture qualitative aspects of the folding process.(191, 192) In this regard we note that unbiased high temperature unfolding simulations of the A β 42 and A β 40 fibril core models sample structures that are similar to those sampled on low energy paths from the calculated free energy surfaces (Figs. 7 and 13 & Figs. 12 and 14). The fact that the unfolding simulations yield observations that are similar to those arising from the lowest energy paths, on the calculated free energy surfaces, argues that the lowest free energy paths are not simply an artifact of the chosen reaction coordinates. Nevertheless, to further assess the significance of these observations, it is important to compare observations arising from the lowest energy path to known experimental observables.

In a recent study, Fawzi et al. performed dark-state exchange saturation transfer (DEST) experiments on A β protofibrils, providing new insights into the dynamics of monomeric A β on the surface of protofibrils (193). While these data suggest that residues in the β 1

and $\beta 2$ strands are essentially equally likely to make direct contacts with the protofibril, measured ^{15}N transverse relaxation rates argue that residues in the $\beta 1$ strand of $\text{A}\beta 40$ are less flexible than residues in the $\beta 2$ strand. Our findings are consistent with these observations in that we find that $\text{A}\beta 40$ binds to the odd end of the fibril through its $\beta 1$ strand while residues in the $\beta 2$ strand remain unstructured and make non-specific contacts with the fibril core (Fig. 11, state b). Indeed, preferential association of the N-termini of an incoming $\text{A}\beta$ peptide was observed in another work examining the energetics of fibril growth.(185) Moreover, a comparison of ^{15}N transverse relaxation rates of $\text{A}\beta 40$ and $\text{A}\beta 42$ further suggests that residues in the $\beta 2$ strand of $\text{A}\beta 40$ are more flexible than residues in the $\beta 2$ strand of $\text{A}\beta 42$.(193) Our data are also consistent with these findings because folding pathways for $\text{A}\beta 42$ contain hairpin structures that are not present in the $\text{A}\beta 40$ folding pathways (Fig. 4, state c). In these structures, residues in the $\beta 2$ strand of $\text{A}\beta 42$ make a series of intramolecular hydrogen bonds that further limit their flexibility. Lastly, a number of mutations have been described that are known to affect the kinetics of fibril formation. For example, one study found that the Flemish mutant A21G decreases the kinetics of fibril extension relative to wild type, while the Dutch mutant E22Q increases it.(194) Another found that the Arctic mutation E22G increased the rate of protofibril formation.(195) It is interesting to note that these mutations and several others cluster in a region of the $\text{A}\beta$ peptide that corresponds to the $\beta 1$ strand of the fibril structures we study (Fig. 1).(195) Since our data suggest that the $\beta 1$ strand associates first and most stably with the odd end of the fibril, it is likely that mutations that increase or decrease the propensity for strand formation in this region would affect fibrillization kinetics.

While the free energy surfaces for $\text{A}\beta 42$ and $\text{A}\beta 40$ fibril elongation share common features, there are significant differences between them. The main difference is that the $\text{A}\beta 42$ monomer undergoes a phase where it essentially “rolls” along the fibril, making contacts with fibril residues that form turns in the structure, before attaching to the odd end, while this behavior is not seen in low energy paths associated with $\text{A}\beta 40$ folding. Moreover, the $\text{A}\beta 42$ folding pathway involves the formation of an S-shaped hairpin

structure (Fig. 4, state c) – a structure that does not occur in the A β 40 fiber elongation pathway. Recent experimental data, in the form of kinetic assays, selective radiolabeling and cell viability experiments, suggest that A β 42 fibrils catalyze the formation of soluble oligomers through a secondary nucleation pathway (76). Rolling of A β 42 monomers on the fibril surface may provide a mechanism for increasing the local concentration of monomeric states. In this sense the fibril surface, particularly regions that form turns in the fibril structure, may provide a secondary nucleation site that enables monomers to self-associate at a higher rate than would be allowed in the surrounding solvent.(76) Moreover, the A β 42 folding pathway involves the formation of an S-shaped structure that has been postulated to exist in A β 42 oligomers(98), a process likely facilitated by the additional 2 hydrophobic C-terminal residues in A β 42. These data are consistent with the notion that soluble oligomers and fibrils share common intermediates with regard to their formation. Since soluble oligomers can induce fibril formation, the presence of the fibril can induce additional fibril deposition via catalyzing the formation of soluble oligomers.(196)

While these results are encouraging, it is important to note that some of these differences between the folding pathways of A β 40 and A β 42 may be due to the fact that the A β 40 starting structure has two filaments while the A β 42 starting structure has one filament (the only available fibril structure for A β 42 at the time that this study was performed). It is difficult to know how our results would generalize if these calculations were performed on A β 42 structures that have multiple filaments. In this vein, we note that after the completion of this work a structural model of A β 40 fibrils was reported that was derived from seeded fibril growth using brain extracts from a patient with Alzheimer's disease.(111) Since this structure is significantly different from the other fibril structures that have been described in the literature, it is not clear how our findings for A β 40 would generalize to these data.

An additional limitation of our study is that the free energy surfaces were computed using an implicit solvent model. Although the solvent model was chosen because it has been shown to yield calculated free energy profiles that are similar to what would be

obtained with explicit solvent (at least for some amyloidogenic peptides)(175, 176), explicit water molecules may play an important role in the kinetics and thermodynamics of A β peptides.(183) Nonetheless, these simulations provide useful insights into the aggregation process and, more importantly, a set of hypotheses that provide fodder for future experiments. For example, our data also suggest that mutations affecting the bend between the β 1 and β 2 strands (residues 23-31) may hinder rolling of incoming monomers and consequently the rate of soluble oligomer formation, albeit it is unclear how this observation generalizes to different fibril morphologies and A β isoforms. In addition, our results argue that longer fibrils would present a larger surface area on which monomers could self-associate, thereby suggesting that the rate of soluble oligomer formation would be increased in the presence of larger the fibrils.

Overall the calculated free energy surfaces provide a new testable hypothesis regarding the mechanism of A β fibril elongation. Indeed, we argue that fibril growth involves the formation of an obligate intermediate, corresponding to a hairpin structure, which forms hydrogen bonds to the odd end of the fibril core. Identification of such an intermediate provides a tunable, druggable pivot in the folding pathway to fibril elongation, and provides a check-point that can be exploited for basic research aiming to elucidate the mechanisms underlying the aggregation process.

All-atom simulations suggest that nucleation of an amyloidogenic peptide proceeds through a helical oligomeric intermediate

Abstract

The aggregation of proteins into amyloid superstructures lies at the heart of several human diseases. Aggregation *in vitro* occurs in two phases: a lag-phase, where nuclei that seed aggregation are formed, and a growth phase, where existing amyloid fibrils are elongated. In this study we explore the early stages of nucleation for an amyloidogenic peptide derived from the non-amyloid- β component of α -synuclein using all atom molecular dynamics simulations in explicit solvent. Simulations are notable for the rapid formation of a helical multimer that unfolds and refolds on the microsecond timescale. During unfolding monomers in the oligomeric unit sample extended states consistent with a β -strand. From these data we derive a model for the nucleation of amyloidogenic peptides where β -sheets form in a stepwise manner via the association of extended monomers arising from unfolding of a dynamic helical oligomer. These data provide new insights into early stages of the nucleation process.

Introduction

Protein aggregation is a wide-spread phenomenon that figures prominently in many human diseases. The most commonly described form of aggregation-related disorders, collectively termed amyloidoses, involve the formation of protein deposits in the form of amyloid fibrils, which are highly repetitive protein superstructures enriched in β -sheets (108, 110, 111, 122, 197). The formation of amyloid aggregates is one of surprising generality, and occurs in a number of otherwise unrelated diseases, including Type 2 diabetes (198, 199), prion disease (200) and various forms of dementia (75). In particular, the aggregation of intrinsically disordered proteins (IDPs) into amyloid fibrils is believed to play a role in the pathogenesis of neurodegenerative diseases. For example, the pathological hallmarks of Alzheimer's and Parkinson's diseases are the presence of amyloid aggregates of Amyloid- β and Tau on the one hand (201, 202), and α -synuclein on the other (105), all of which have been shown to be intrinsically disordered in the monomeric state (16, 135, 203). Several studies suggest that the primary toxic species in these diseases are soluble oligomeric aggregates rather than amyloid fibrils (8, 15, 79). While the precise mechanism of toxicity has yet to be elucidated, it is evident that aggregation of otherwise disordered, predominantly monomeric proteins plays an important role, and the process appears to be of great generality across different diseases (75). As a result there is considerable interest in understanding the aggregation process.

Amyloid formation occurs in two phases, an extended lag phase in which little detectable fibrillar material is formed but during which nucleation occurs, followed by a rapid growth phase in which amyloid fibrils are produced until the surrounding monomeric pool is depleted (204). Thus far, the end products of this process, amyloid fibrils, have proven to be the most amenable to structural characterization due to their highly repetitive and stable nature (108, 109, 197, 205). The events occurring during the earlier phases, however, remain a mystery. While it is likely that the nucleation process ends in protofibrillar species that can form the basis for fibril growth, the structural

events leading to the formation of β -strand rich protofibrils from individual monomers are unclear. This is in large part due to the technical difficulties associated with characterizing metastable species that are likely to occur on timescales that are too short to be tractable by traditional methods for structure determination. In the present study we present a 34.6 μ s all-atom simulation of a system consisting of amyloidogenic peptide monomers at a concentration known to promote the formation of amyloid fibrils. Our goal is to gain insight into the earliest events in the nucleation process.

Results and Discussion

We chose to simulate the non-amyloid- β component region, NAC(8-18), of α -synuclein, an 11 residue hydrophobic segment, having the sequence GAVVTGVTAVA, which has been shown to be the minimum fragment of α -synuclein to induce toxicity and aggregate into amyloid fibrils *in vitro* (154). Circular dichroism experiments indicate that this peptide natively adopts a random coil structure (206). Two sets of simulations were performed. We first performed simulations of a single peptide in TIP3P water to study the conformational preferences of the peptide alone. In a second set of simulations we placed four copies of the peptide into a TIP3P cubic solvent box approximately 80 \AA across, corresponding to a concentration of 5mM – a concentration that is 5 times larger than what is needed for fibers to form *in vitro* (154). These latter simulations were chosen to see how conformational preferences of this amyloidogenic peptide changes under conditions that favor fibril formation. In both cases salt was included to achieve a concentration of 150mM NaCl. Both systems were equilibrated for 2ns before being loaded onto a 512-node Anton special-purpose supercomputer, designed specifically for molecular dynamics simulations, for production dynamics (207).

Simulations of the isolated NAC(8-18) peptide reveal that the peptide samples conformations that are enriched in both helical and strand content (Fig. 1). The average helical and strand contents of the monomer are 0.44 ± 0.35 and 0.11 ± 0.13 , respectively (Fig. 1). In addition to the fully unfolded state (U, fractional secondary

structure content < 0.2), two types of helical conformations are sampled during the trajectory – a partially helical conformation (P, helical content 0.2-0.7), and a fully helical state (H, helical content > 0.7). State P is sampled most frequently and accounts for 43.3% of structures sampled. State H occurs in 25.6% of structures sampled from the trajectory, and state U occurs in 12.5% of structures (Fig.1). In addition, the system occasionally samples more extended states: structures with a strand content of more than 0.5 (E, for extended) are sampled 1.2% of the time, suggesting that extended states are accessible to the monomer but are less favorable than the helical conformations (Fig. 1).

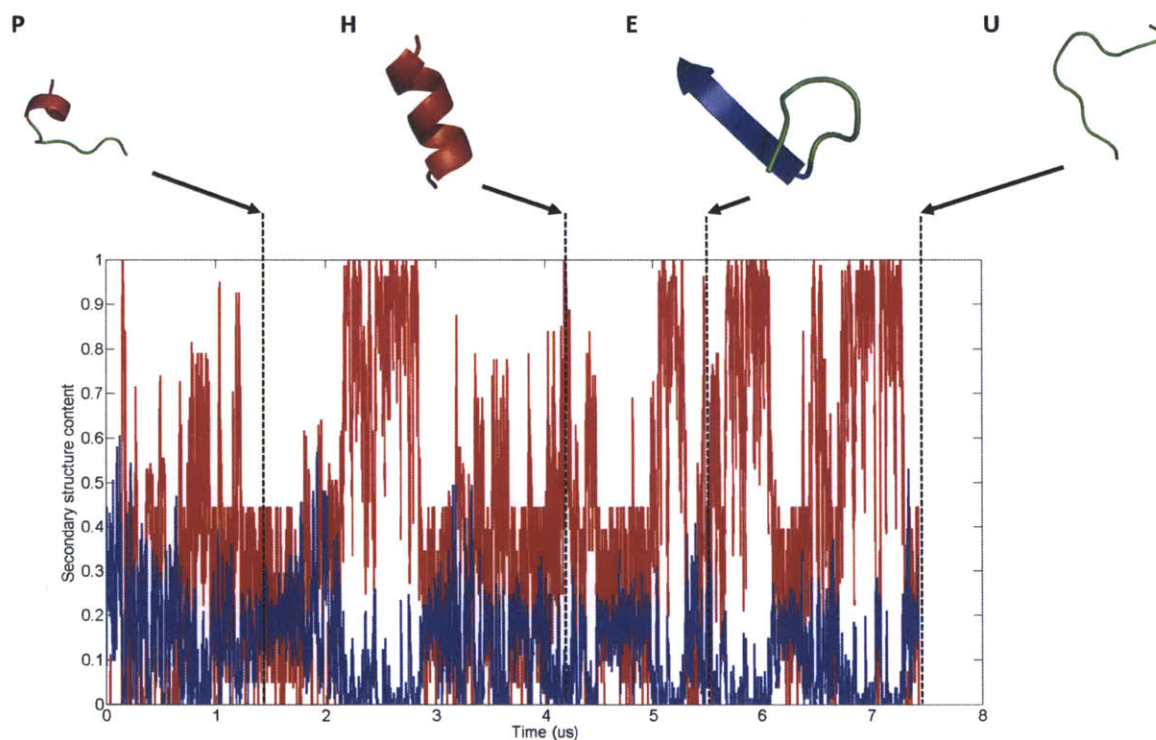


Figure 1. Helical and strand contents (red and blue, respectively) of monomeric NAC(8-18) over the course of the 7.6 μ s production run, as determined by DSSP. Also shown are representative structures for the different states discussed in the text. Helices are represented in red and strands in blue.

Simulations of the high concentration system display markedly different behavior. In these simulations monomers assemble into helical structures within $3\mu\text{s}$ and a helical tetramer is formed by $5\mu\text{s}$ (Figs. 2A and B). The average helical and strand content of the system consisting of four-monomers is 0.87 ± 0.24 and 0.02 ± 0.07 , respectively. These data suggest that at high concentrations helical states are favored, relative to the isolated peptide simulations.

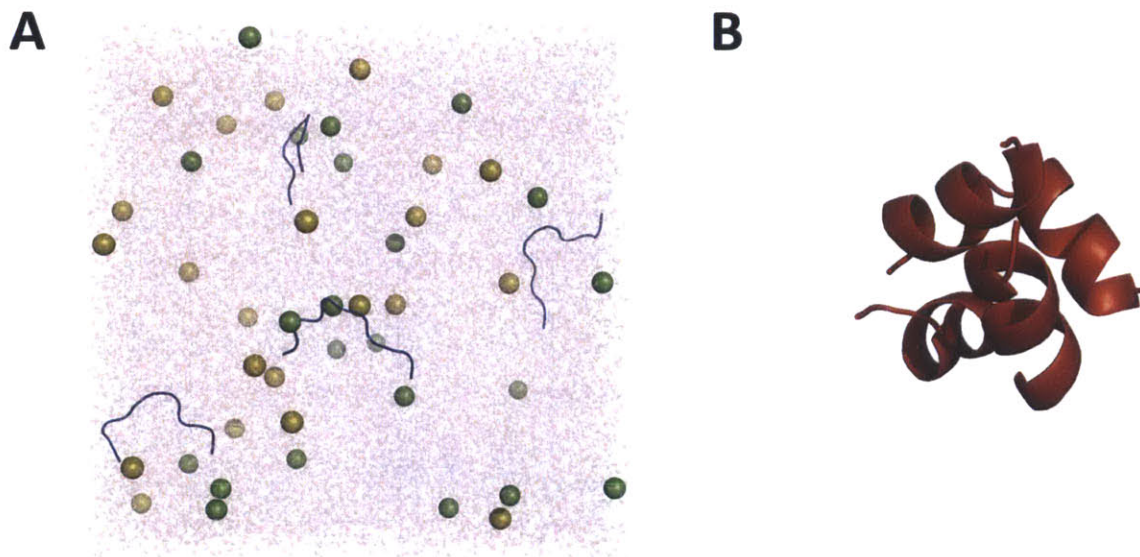


Figure 2. (A) Initial state of the four monomer system at the beginning of the production run. Monomers are represented by cartoons, Na^+ and Cl^- ions are shown as yellow and green spheres, respectively, and bulk TIP3P solvent is shown in transparent. (B) Snapshot of the helical tetrameric state, taken after $5\mu\text{s}$ of simulation time.

Once the helical tetramer is formed, stochastic unfolding and refolding of monomers in the tetramer occurs on the microsecond time scale. Over a total of $34.6\mu\text{s}$, several distinct partially unfolded states can be identified: the folded helical tetramer, H_n , occurring in 57.2% of structures sampled from the trajectory; structures containing one partially unfolded monomer ($\text{H}_{n-1}\text{-P}$, occurring in 33.4% of structures) or fully unfolded

monomer (H_{n-1} -U, occurring in 6.8% of structures); and structures containing a monomer that adopts an extended conformation, (H_{n-1} -E), occurring in 0.8% of structures (Fig. 3).

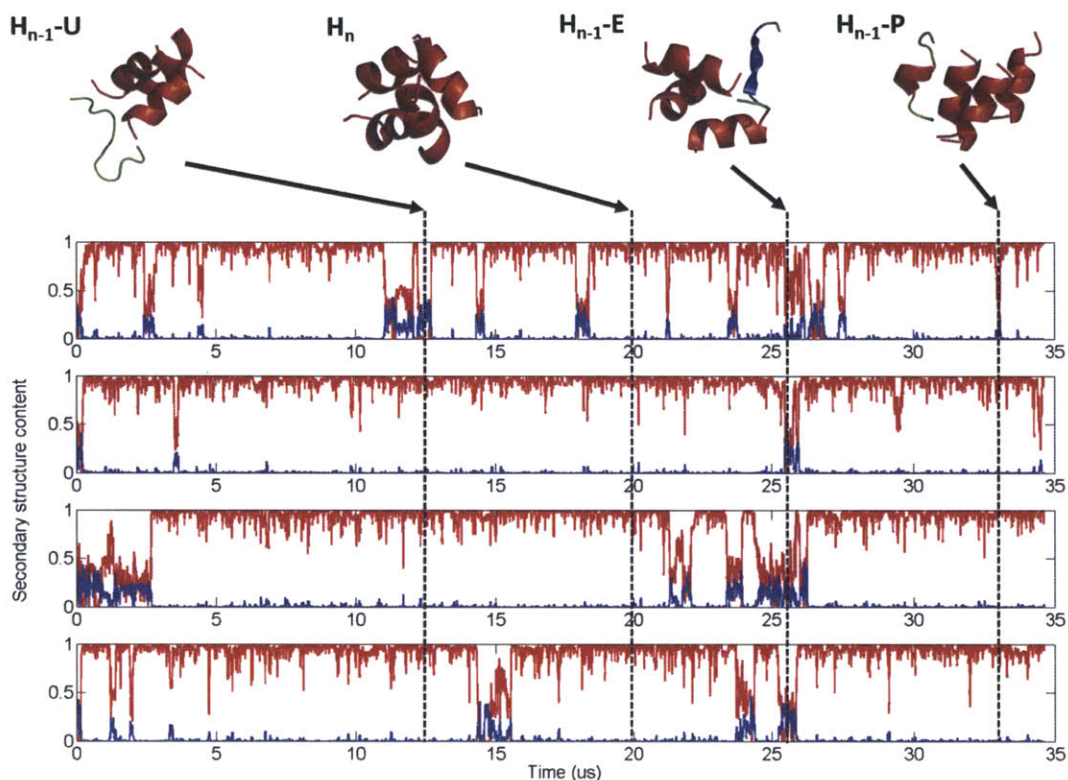


Figure 3. Secondary structure contents of each of the four NAC(8-18) monomers over the course of the 34.6 μ s production run, as determined by DSSP. Helical content is shown in red, and strand content in blue. Also shown are the dominant states discussed in the text (whose secondary structural elements correspond to their DSSP assignments): H_n , H_{n-1} -P, H_{n-1} -U and H_{n-1} -E. The arrows indicate both the time points from which these species were extracted.

Based on these observations we propose a model for the stochastic events leading to the nucleation of NAC(8-18) and the formation of protofibrils (Fig. 4). Freely-floating monomers fluctuate between states U, P and H. Random encounter of monomers leads

to the formation of dynamic helical aggregates, H_n , composed of n monomers in state H. Stochastic unfolding of helical monomers to states P or U occurs while the monomers themselves remain in contact with the helical oligomer (e.g., states H_{n-1} -P, H_{n-1} -U, H_{n-1} -E) – a mechanism that allows monomers to sample additional states while remaining in close proximity to other monomers. Most of the time, the structure containing an extended monomer, H_{n-1} -E, folds back into state H_n (Fig. 3). However, in the extended state a monomer is available to make backbone hydrogen bonds with other nearby monomers that are in the helical oligomer. In this model conversion from an entirely helical oligomer H_n to a strand-based oligomer S_n occurs in a stepwise fashion. Stochastic unfolding of individual helices occurs one by one, allowing their backbone donors and acceptors to become exposed and available for the formation of inter-molecular backbone hydrogen bonds (Fig. 4).

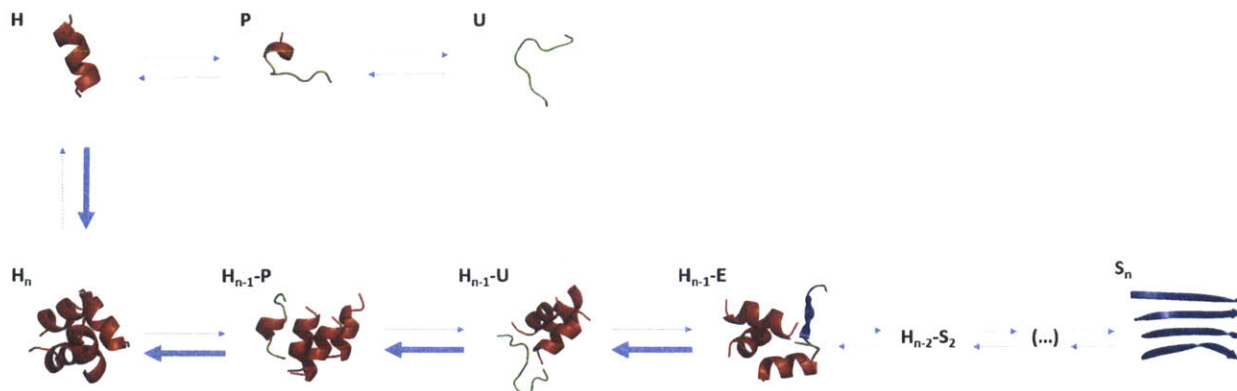


Figure 4. Proposed model for the early stages of aggregation of the NAC(8-18) peptide. States H , P , U , H_n , H_{n-1} -P, H_{n-1} -U and H_{n-1} -E are described in the text. H_{n-2} - S_2 corresponds to a state in which two adjacent monomers are extended, and form a sheet through backbone hydrogen-bonds. S_n corresponds to a state in which all monomers within the oligomer nucleus are extended and arranged in a sheet.

This offers an explanation for the slow nature of nucleation, which is dependent on stochastic process requiring the crossing of multiple, reversible check-points before reaching a (meta)stable nucleus S_n that can form the basis for further aggregation by acting as a template.

While the literature on the early stages in the nucleation process is sparse, early formation of helical structures prior to the formation of beta-strand rich fibrils has been experimentally described for some systems. The myostatin precursor protein, for example, whose aggregation has been linked to sporadic inclusion body myositis, showed a distinctly helical CD spectrum for early-stage soluble aggregates, which turned into a spectrum indicating a sample rich in strands after overnight incubation (208). Similarly, a class of synthetic tri- and hexapeptides, rationally designed to simulate fibril assembly, also progressed through helical oligomeric intermediates. These helical species appeared in a concentration-dependent manner starting at roughly 1mM and turned into strand-rich aggregates over time (209). Secondary structure content was determined by circular dichroism spectroscopy and the concentration-dependence of aggregate formation was inferred from X-ray fiber diffraction studies (209). Our model is also consistent with observations arising from a simulations arising from a simplified tube-like model of a peptide whose native state is helical but that is known to form amyloids (210).

The current study argues that the formation of helical aggregates is an early step in the formation of amyloid structures even if the amino-acid sequence in question does not have a strong preference for helical structure. The formation of a dynamic helical oligomer occurs rapidly, followed by stochastic unfolding and refolding of its constituent monomers on the microsecond timescale. Stochastic unfolding facilitates the sampling of extended monomeric structures that can, in principle, self-associate. These observations form the basis for a physically viable model that we hope will fodder additional studies of the early stages in the nucleation process.

Materials and Methods

An all-atom model of the NAC(8-18) peptide, with sequence GAVVTGVTAVA, was created and minimized in CHARMM. The N-terminus was capped with an acetyl group and the C-terminus was capped with an N-methylamide group, in order to mimic the behavior of NAC(8-18) fragments in the context of the full-length NACP or even full-length α -synuclein and avoid spurious self-association occurring due to do interactions between charged termini. The monomer system was solvated in a cubic TIP3P lattice and 150mM NaCl. The four monomer system was created by placing the monomers in a plane, with their centers of mass translated 25Å in separate directions from the starting structure before adding TIP3P water and 150mM NaCl. Both systems were then equilibrated for 2ns in CHARMM, before being loaded onto a 512 node Anton special-purpose machine for the production runs. Production runs were performed in the CHARMM22 forcefield (211) with periodic boundary conditions, using 2.5fs timesteps and the Multigrator integrator at constant pressure and temperature. Anton uses the k -space Gaussian split Ewald method to compute electrostatic interactions: short-range electrostatics were computed directly using Coulomb's law, while long-range interactions were computed on a particle mesh using an Ewald sum (207). Structures were sampled from the trajectory every 240ps. Secondary structure contents were computed using the DSSP program (147).

Conclusions and future directions

This dissertation details three separate studies of aggregation in IDPs related to neurodegeneration. In all three studies, we observe a similar phenomenon: a polypeptide chain that is not prone to adopting ordered, stable structure in its monomeric state can nonetheless adopt it when in the proximity of a binding partner, which in the aggregation process is usually at least one other protein of the same type. Often, these neighboring proteins will need to be in a particular conformation in order to facilitate association with the incoming monomer. In the first study, we showed that NMR data suggests the existence of a small amount of multimeric conformations of alpha-synuclein containing extensive secondary structure content, a property that is not observed in the free monomeric protein. In the second study, we showed that the relatively flat free energy landscape of a disordered monomer of A β is transformed through proximity to an amyloid fibril in such a manner as to then have a well-defined global free energy minimum corresponding to the folded, bound state. Finally, we showed in the third study that while the minimum aggregating fragment of alpha-synuclein only adopts a small amount of secondary structure in isolation, it becomes much more prone to adopting a conformation rich in secondary structure when in the presence of other monomers at a concentration known to lead to aggregation. These findings highlight the fact that the free energy landscape of a protein (and particularly an IDP) can only be defined for a particular physico-chemical environment. If this environment changes, the energy landscape changes, and therefore so does the set of

thermally accessible conformations and their relative frequencies. This concept should not be unfamiliar to people who consider prion proteins, which operate on this very principle, and indeed there is evidence to suggest that alpha-synuclein behaves like a prion by inducing aggregation in neurons through the influx of aggregation-prone conformations from their neighbors. However, most studies of IDPs are performed *in vitro* in relatively homogeneous samples, so it is likely that the observed energy landscapes differ significantly from the crowded environment of the cell. One could argue that this renders the concept of a free energy landscape rather ill-defined for an IDP in anything but the most dilute and homogeneous sample, but this difficulty can in part be overcome by appropriate consideration of specific, well-defined environments the protein might experience in its biological lifetime, such as the presence of lipids, neighboring aggregates or binding partners of an entirely different type, and restricting the discourse of the protein's energy landscape to that particular environment. In addition, improvements of *in vivo* experimental techniques, such as in-cell NMR (137), will allow us to hone in on clinically relevant environments.

While it is the author's opinion that these studies show that useful insights can be generated by considering the energy landscape of IDPs in different situations, either by leveraging experimental data with computation, or with computational modeling alone, it is clear that we are far from a complete description of the aggregation pathway of even a single IDP. Models of these pathways and their constituent intermediates remain highly speculative and coarse-grained in nature, and experiments are often inconclusive. This is not for want of trying, but rather is an inevitable consequence of the limitations in the state-of-the-art in both experimentation and computation relative to the magnitude of the challenge of characterizing an IDP.

Difficulties in studying these systems arise from the fact that IDPs, like all proteins, are too small to observe directly and in addition, change shape very rapidly, adopt a very large set of structurally dissimilar states, and aggregate on very slow timescales. Thus, improvements in our ability to characterize IDPs will have to rest on our ability to obtain,

on the one hand, very high resolution measurements in both time and space in order to capture the conformational propensities of individual monomers, and to bridge the gap with the slow timescales of aggregation on the other. As far as spatio-temporal resolution is concerned, the advent of time-resolved, single-molecule techniques such as single-molecule FRET are promising in their ability to observe the backbone dynamics of individual IDPs *in vitro* (117), and maybe even *in cellulo*, though the quality of the structural information these techniques could offer will depend on the extent to which the covalent attachment of fluorophores at different sites on the polypeptide chain affects the underlying energy landscape of the protein. While we know that the wavelength of visible light places a hard limitation on what can be observed directly with our eyes, it is not impossible that alternate forms of microscopy may one day allow the direct observation of conformational changes in an individual protein, as evidenced by the increasing development of nanometer-level resolution single-particle tracking techniques (212). As experimental data improve in quality and resolution, they will become less and less degenerate, and describe the protein's conformation more uniquely at specific time points. This could allow for the construction of higher resolution conformational ensembles with lower uncertainty using the same formalism described in the first study. These Bayesian weighted ensembles could further be strengthened by inclusion of previously obtained structural data on the same system in the same conditions into the prior distribution for the weights (in the first study, the prior was chosen to be an uninformative prior, reflecting a state of ignorance about the underlying weights).

Bridging the gap between the timescales of individual monomer dynamics and aggregation events will likely involve advances in computing as well as experiment. Increasing molecular dynamics forcefield accuracy with better benchmarking to experimental IDP systems in combination with improvements in application-specific computing hardware will allow for atomistic investigations of molecular events at ever increasing timescales, but the fact that the aggregation events pertaining to neurodegeneration occur on the timescale of a human life (as evidenced by the fact that

neurons are more susceptible to housing aggregated forms than other cells due to their longevity) casts a looming shadow over these efforts by highlighting the number of orders of magnitude separating the timescales of these two regimes. The extent of this timescale separation can presumably be reduced through appropriate tweaking of the experimental conditions, much as is done in amyloid aggregation assays (e.g. by artificially increasing the concentration of protein *in vitro*), and through an artful distillation of the essential features of the system into simpler, more tractable computational models. Ultimately, however, we need not be disheartened by the extent of the difficulty in studying IDPs, as they showcase the most sophisticated, naturally-occurring nanotechnology in existence in the known natural world, and are likely to continue to fascinate and bewilder us for many years to come.

Appendix

A1 - Definition of f_β , a reaction coordinate quantifying strand content

We define the fraction of strand content in a polypeptide chain composed of N amino acids, computed as a function of its backbone dihedral angles, as the average strand content of each residue:

$$f_\beta(\phi_1, \psi_1, \phi_2, \psi_2, \dots, \phi_N, \psi_N) = \frac{1}{N-2} \sum_{i=2}^{N-1} f_\beta^{(i)}(\phi_i, \psi_i) \quad (1.1)$$

where the sum omits the contents of the first and last residue, i.e. residues that do not have both ϕ and Ψ angles that are well defined. The content of an individual residue is defined according to Vitalis *et al.* (12):

$$f_{\beta}^{(i)}(\phi_i, \psi_i) = \begin{cases} 1 & \text{if } (\phi_i - \phi_{\beta})^2 + (\psi_i - \psi_{\beta})^2 < r_{\beta}^2 \\ \exp(-\tau_{\beta} D_{(i)}^2) & \text{otherwise} \end{cases}, \quad (1.2)$$

where

$$D_{(i)}^2 = \left(\sqrt{(\phi_i - \phi_{\beta})^2 + (\psi_i - \psi_{\beta})^2} - r_{\beta} \right)^2, \quad (1.3)$$

the square of the Euclidean distance between (ϕ_i, ψ_i) and the boundary of the basin. Equation (1.2) defines a circular basin in Ramachandran space, with center $(\phi_{\beta}, \psi_{\beta})$ and radius r_{β} , in which the strand content for that residue has a value of 1. Outside this basin, the strand content decays exponentially with decay constant τ_{β} . Note that due to the periodic nature of dihedral angles, any distance in Ramachandran space, including angle differences and $D_{(i)}^2$, is assumed to be the minimum possible distance in that space. The parameters used in works described in this document were $(\phi_{\beta}, \psi_{\beta}) = (-152.0^{\circ}, 142.0^{\circ})$, $r_{\beta} = 62.0^{\circ}$ and $\tau_{\beta} = 0.0029 \text{deg}^{-2}$.

The circular basin for strand content does not assign a ‘preferred’ set of dihedral angles for strand content, but rather considers any angles within the basin to be equal in this regard. This is based on the understanding that the probability distribution $g(\vec{\phi}, \vec{\psi})$ of backbone dihedral angles across protein structures deposited in the Protein Data Bank (PDB) (58) is similarly flat in this region. In contrast, inspection of the helical region of $g(\vec{\phi}, \vec{\psi})$ shows a clear peak, and a non-zero skewness. Thus, a definition for the fraction of helical content, f_{α} , based on dihedral angles would better fit the helical region of the probability distribution across the PDB with a function that has a clear maximum, such as a 2-dimensional Gaussian. If one wished to capture the skewness observed in the helical region of $g(\vec{\phi}, \vec{\psi})$, however, a distribution with a non-zero third moment would be required.

A2 - Derivation of molecular dynamics forces arising from introducing umbrella potentials in f_β

Umbrella sampling is performed by adding ‘umbrella potentials’ to the potential energy being used to evaluate your system (e.g. a molecular dynamics potential): $U_{total} = U_{system} + U_{umbrella}$. An umbrella potential will typically involve applying a harmonic restraint to a reaction coordinate of choice ξ , where the harmonic well is centered about a specific value ξ_0 . A potential of mean force (PMF), or free energy surface (FES), can then be computed from the umbrella sampling windows, to obtain the free energy of the system as a function of the reaction coordinate, $G(\xi)$. When constructing a FES as a function of f_β , an umbrella potential would take the form $U_{f_\beta} = k_{f_\beta} (f_\beta - f_0)^2$. If the umbrella sampling is to be performed using molecular dynamics (MD), it is necessary to differentiate the umbrella potential in order to compute the forces acting on the system. In the case of simple reaction coordinates such as inter-atomic distances, molecular dynamics packages will frequently contain pre-packaged functions that can be used to introduce restraints (e.g. *RESDistance* in CHARMM (171)), and the forces will be added to the pre-existing forcefield automatically. In the case of a more elaborate reaction coordinate such as f_β , the forces need to be hardcoded into the molecular dynamics forcefield. This subsection will detail the calculation of these forces and how they can be applied to the already existing CHARMM forcefield.

The forces acting on a given atom at position \vec{x} are calculated from $F_{total} = -\nabla U_{total}$. After introducing an umbrella potential to the CHARMM forcefield, our total potential energy is $U_{total} = U_{CHARMM} + U_{f_\beta}$. Since the gradient ∇ is a linear operator, it is clear that

$$F_{total} = F_{CHARMM} - \nabla U_{f_\beta}. \quad (2.1)$$

In other words, if we can compute the negative gradient of the umbrella potential, $-\nabla U_{f_\beta}$, for a given atom, it can simply be added to the force already acting on that atom by the CHARMM forcefield. We find that

$$\begin{aligned}
-\nabla U_{f_\beta} &= -\frac{dU_{f_\beta}}{d\bar{x}} \\
&= -2k_{f_\beta} (f_\beta - f_0) \frac{df_\beta}{d\bar{x}} \\
&= -\frac{2k_{f_\beta}}{N} (f_\beta - f_0) \sum_{i=1}^N \frac{df_\beta^{(i)}}{d\bar{x}}
\end{aligned} \tag{2.2}$$

where

$$\frac{df_\beta^{(i)}}{d\bar{x}} = \begin{cases} 0 & \text{if } D_{(i)} < 0 \\ -\tau_\beta \exp(-\tau_\beta D_{(i)}^2) \frac{dD_{(i)}^2}{d\bar{x}} & \end{cases} \tag{2.3}$$

and

$$\frac{dD_{(i)}^2}{d\bar{x}} = \frac{2\sqrt{D_{(i)}^2}}{\sqrt{(\phi_i - \phi_\beta)^2 + (\psi_i - \psi_\beta)^2}} \left[(\phi_i - \phi_\beta) \frac{\partial \phi_i}{\partial \bar{x}} + (\psi_i - \psi_\beta) \frac{\partial \psi_i}{\partial \bar{x}} \right] \tag{2.4}$$

Thus, after application of the chain rule, we are left with the gradients of each backbone

dihedral angle i with respect to Cartesian coordinates, $\frac{\partial \phi_i}{\partial \bar{x}}$ and $\frac{\partial \psi_i}{\partial \bar{x}}$. These can be

computed using the procedure for computing a dihedral angle derivative $\frac{\partial \phi}{\partial \bar{x}}$ first

described by Arnaud Blondel and Martin Karplus in 1996 (213), detailed below.

A proper dihedral (or torsion) angle is defined for four connected atoms with

coordinates r_i, r_j, r_k and r_l . We define a first set of intermediate vectors

$$F = r_i - r_j, \quad G = r_j - r_k, \quad H = r_l - r_k, \tag{2.5}$$

from which we can construct a second set of intermediate vectors as

$$A = F \otimes G, \quad B = H \otimes G \tag{2.6}$$

The dihedral angle of interest, φ , can be defined by the angle between A and B as

$$\cos\varphi = \frac{A \cdot B}{|A||B|} \quad (2.7)$$

$$\sin\varphi = \frac{B \otimes A \cdot G}{|A||B||G|}. \quad (2.8)$$

We can then calculate the derivative of φ by using the derivative of $\cos(\varphi)$ as follows:

$$\frac{\partial\varphi}{\partial r} = \frac{\partial\varphi}{\partial\cos\varphi} \cdot \frac{\partial\cos\varphi}{\partial r}. \quad (2.9)$$

Since

$$\frac{\partial\varphi}{\partial\cos\varphi} = \frac{-1}{\sin\varphi}, \quad (2.10)$$

we obtain

$$\frac{\partial\varphi}{\partial r} = \frac{-1}{\sin\varphi} \cdot \frac{\partial\cos\varphi}{\partial r}. \quad (2.11)$$

In order to evaluate $\frac{\partial\varphi}{\partial r}$ from (2.11), we apply the chain rule sequentially to $\frac{\partial\cos\varphi}{\partial r}$ to

break it down to derivatives with respect to the second intermediate vectors A and B , followed by derivatives with respect to F , G and H . Using the definition of $\cos(\varphi)$ in (2.7), we find:

$$\frac{\partial\cos\varphi}{\partial A} = \frac{B}{|A||B|} + \frac{A \cdot B}{|B|} \cdot \frac{\partial(1/|A|)}{\partial A}. \quad (2.12)$$

Using the identity

$$\frac{\partial |A|}{\partial A} = \frac{|A|}{A}, \quad (2.13)$$

we obtain

$$\frac{\partial \cos \varphi}{\partial A} = \frac{1}{|A|^3 |B|} (A^2 B - (A \cdot B) A). \quad (2.14)$$

In addition, using the double cross product formula, we have the identity:

$$A \otimes (B \otimes A) = A^2 B - (A \cdot B) A. \quad (2.15)$$

Plugging (2.15) into (2.14) and comparing with (2.8), using the fact that G is orthogonal to both A and B , we obtain:

$$\frac{\partial \cos \varphi}{\partial A} = \frac{1}{A^2 |G|} \sin(\varphi) A \otimes G. \quad (2.16)$$

We can then replace r with A in equation (2.11), and substituting for $\frac{\partial \cos(\varphi)}{\partial A}$ in (2.16)

we find:

$$\frac{\partial \varphi}{\partial A} = \frac{1}{A^2 |G|} G \otimes A. \quad (2.17)$$

This step involves $\frac{\sin(\varphi)}{\sin(\varphi)}$, so (2.17) is true for all φ with $\sin(\varphi) \neq 0$. However, (2.17) is independent of vectors H and B and thus is independent of φ , which means that it also extends to the case $\sin(\varphi) = 0$. This lack of discontinuity in the torsional derivative is the motivation behind this method for computing forces relating to a torsional potential (213). Appealing to the vector definitions in (2.5) and (2.6), interchanging $i \leftrightarrow l$ and $j \leftrightarrow k$ leads to the symmetries $F \leftrightarrow H, G \leftrightarrow -G, A \leftrightarrow -B$ and $\varphi \leftrightarrow -\varphi$. Using this fact, we have the corresponding equation:

$$\frac{\partial \varphi}{\partial B} = \frac{1}{B^2 |G|} B \otimes G. \quad (2.18)$$

Next, we determine the derivatives with respect to F , G and H through application of the chain rule. Using (2.6), we find:

$$\frac{\partial A}{\partial F} = \frac{\partial(F \otimes G)}{\partial F} = I \otimes G, \quad (2.19)$$

where I is the identity matrix and $I \otimes G$ is defined such that $I \otimes G \cdot V = V \otimes G$, where V is any vector. Using equations (2.17) and (2.19) and the chain rule, we obtain:

$$\frac{\partial \varphi}{\partial F} = \frac{\partial \varphi}{\partial A} \frac{\partial A}{\partial F} = \frac{1}{A^2 |G|} (G \otimes A) \cdot I \otimes G. \quad (2.20)$$

The transpose sign on $\frac{\partial \varphi}{\partial A}$ is used to emphasize the importance of the order of the terms in the chain rule because $\frac{\partial A}{\partial F}$ is antisymmetric. Evaluating (2.20) using the definition of $I \otimes G$ yields:

$$\frac{\partial \varphi}{\partial F} = -\frac{1}{A^2 |G|} (G \otimes A) \otimes G. \quad (2.21)$$

The minus sign occurs due to the antisymmetry of (2.20). With equation (2.15) and the identity $A \cdot G = 0$, (2.21) reduces to:

$$\frac{\partial \varphi}{\partial F} = -\frac{|G|}{A^2} A. \quad (2.22)$$

By symmetry, we have

$$\frac{\partial \varphi}{\partial H} = \frac{|G|}{B^2} B. \quad (2.23)$$

Using the chain rule, the equivalent of (2.19) to obtain $\frac{\partial A}{\partial G} = F \otimes I$ and $\frac{\partial B}{\partial G} = H \otimes I$ and

their antisymmetry, we find that

$$\begin{aligned} \frac{\partial \varphi}{\partial G} &= \frac{\partial \varphi}{\partial A} \frac{\partial A}{\partial G} + \frac{\partial \varphi}{\partial B} \frac{\partial B}{\partial G} \\ &= \frac{1}{A^2 |G|} (G \otimes A) \otimes F - \frac{1}{B^2 |G|} (G \otimes B) \otimes H \end{aligned} \quad (2.24)$$

We replace A and B by (2.6) and use (2.15) to obtain:

$$\frac{\partial \varphi}{\partial G} = \frac{(G^2 F - (F \cdot G)G) \otimes F}{A^2 |G|} - \frac{(G^2 H - (H \cdot G)G) \otimes H}{B^2 |G|}. \quad (2.25)$$

Using $F \otimes F = 0$, $H \otimes H = 0$ and (2.6), we obtain:

$$\frac{\partial \varphi}{\partial G} = \frac{(F \cdot G)}{A^2 |G|} A - \frac{(H \cdot G)}{B^2 |G|} B. \quad (2.26)$$

Finally, we evaluate the derivatives of F , G and H with respect to r , to obtain the final expressions for $\frac{\partial \varphi}{\partial r}$. The non-zero terms can be seen from (2.5):

$$\frac{\partial F}{\partial r_i} = I, \quad \frac{\partial F}{\partial r_j} = -I, \quad \frac{\partial G}{\partial r_j} = I, \quad \frac{\partial G}{\partial r_k} = -I, \quad \frac{\partial H}{\partial r_k} = -I \quad \text{and} \quad \frac{\partial H}{\partial r_l} = I. \quad (2.27)$$

Combining all the above results, we obtain:

$$\frac{\partial \varphi}{\partial r_i} = -\frac{|G|}{A^2} A \quad (2.28)$$

$$\frac{\partial \varphi}{\partial r_j} = \frac{|G|}{A^2} A + \frac{(F \cdot G)}{A^2 |G|} A - \frac{(H \cdot G)}{B^2 |G|} B \quad (2.29)$$

$$\frac{\partial \phi}{\partial r_k} = \frac{(H \cdot G)}{B^2 |G|} B - \frac{(F \cdot G)}{A^2 |G|} A - \frac{|G|}{B^2} B \quad (2.30)$$

$$\frac{\partial \phi}{\partial r_l} = \frac{|G|}{B^2} B \quad (2.31)$$

A3 - Implementation in CHARMM

These equations (2.28-2.31) can be plugged into (2.4) to complete the overall derivative, by replacing for $-\nabla U_{f_\beta}$, where (i, j, k, l) correspond to $(C_{m-1}, N_m, CA_m, C_m)$ for the ϕ angle and $(N_m, CA_m, C_m, N_{m+1})$ for the ψ angle of residue m 's backbone, respectively. For each backbone atom in residues that have well-defined ϕ and ψ angles, these expressions can be added to the force arrays in the already existing CHARMM forcefield. These arrays describe the forces acting on each atom in the system. In addition, the total potential energy must be updated by adding the potential energy arising from the additional calculations to the already existing CHARMM potential energy. Both of these tasks are accomplished by populating the otherwise blank USERE subroutine in the source/charmm/usersb.src source file of the CHARMM source code. This subroutine allows the user to specify additional calculations that will be performed at every timestep of the simulation. Its inputs can be modified to include the force arrays or any other global CHARMM variable, provided calls to this subroutine are appropriately updated throughout the CHARMM source code.

In the code, one specifies a minimum f_0 for the umbrella potential $U_{f_\beta} = k_{f_\beta} (f_\beta - f_0)^2$. The whole CHARMM package can then be recompiled to yield an executable that will implement the desired umbrella potential. In order to perform umbrella sampling, the user would have to recompile separate executables for each increment in f_0 between 0 and 1 that one wishes to consider. The above method can be implemented across the entire polypeptide chain (i.e. from the second to the penultimate residue), or for specific

regions of it, as desired. It goes without saying that the number of residues considered should be an integer multiple of the number of increments in increments f_0 for the reaction coordinate to be interpretable.

A4 - Theoretical foundations of umbrella sampling

In this section we outline the motivation behind the umbrella sampling method for free energy calculations. The goal is to compute the free energy of the system under study, $G(\xi)$, as a function of a chosen reaction coordinate ξ . Umbrella sampling provides a formalism for combining a set of biased probability distributions, each generated by introducing an umbrella potential, typically of the form $U_{umbrella} = k(\xi - \xi_0)^2$ as described in the previous sections, that applies a harmonic restraint to the system ξ about a given value of ξ_0 . To avoid confusion with the total potential energy of the system U , we will use the notation $V_{umbrella}$ for an umbrella potential throughout the remainder of this section. The degrees of freedom of our system form a set $D = \{x_1, \dots, x_{3N}\}$ representing the Cartesian coordinates of the atoms in the system. A transformation can be performed on the set D such that the system is described according to the degrees of freedom $D' = \{\xi, y_1, \dots, y_{3N-1}\}$, where ξ is the reaction coordinate and y_1, \dots, y_{3N-1} are transformed degrees of freedom which are complicated functions of the elements of D . We can therefore express the probability of observing a value of ξ by integrating out all other degrees of freedom in the set D' , such that

$$p(\xi) = \frac{\int \exp(-\beta U(\xi, y_1, \dots, y_{3N-1})) dy_1 \dots dy_{3N-1}}{\int \exp(-\beta U(\xi, y_1, \dots, y_{3N-1})) d\xi dy_1 \dots dy_{3N-1}} \quad (4.1)$$

$$= \frac{\exp(-\beta U(\xi))}{Z},$$

where $\beta = 1/k_B T$, the reciprocal of the thermal energy, $U(\xi, y_1, \dots, y_{3N-1})$ is the total potential energy of the system in a given configuration defined by the elements of D' , Z

is the canonical partition function for our system and $U(\xi)$ the internal energy of the system in window i as a function of ξ . In a given window i , we have a biasing potential $V_i(\xi)$ and an unbiased probability $p_i(\xi)$ of observing ξ in that window. The partition function Z is a constant statistical mechanical property of the system, so we can write this probability as

$$p_i(\xi) = \frac{\exp(-\beta U_i(\xi))}{Z} \quad (4.2)$$

Multiplying by one twice, we can write (4.2) as

$$p_i(\xi) = \frac{\exp(-\beta U_i(\xi)) \cdot \exp(-\beta V_i(\xi)) \cdot \exp(\beta V_i(\xi))}{Z} \cdot \frac{Z_i}{Z_i}, \quad (4.3)$$

where we define Z_i^* to be the partition of the biased system, according to

$$Z_i^* = \int \exp(-\beta(U_i(\xi) + V_i(\xi))) d\xi dy_1 \dots dy_{3N-1}. \quad (4.4)$$

Equation (4.3) can be rewritten as

$$p_i(\xi) = \frac{\exp(-\beta(U_i(\xi) + V_i(\xi)))}{Z_i^*} \exp(\beta V_i(\xi)) \cdot \frac{Z_i}{Z}, \quad (4.5)$$

where the left-most term of the right-hand side is the biased probability $p_i^*(\xi)$. If we write the partition functions explicitly, it becomes clear that

$$\begin{aligned}
\frac{Z_i}{Z} &= \frac{\int \exp\left(-\beta(U_i(\xi, y_1, \dots, y_{3N-1}) + V_i(\xi))\right) d\xi dy_1 \dots dy_{3N-1}}{\int \exp\left(-\beta U_i(\xi, y_1, \dots, y_{3N-1})\right) d\xi dy_1 \dots dy_{3N-1}} \\
&= \frac{\int \exp\left(-\beta U_i(\xi, y_1, \dots, y_{3N-1})\right) \exp\left(-\beta V_i(\xi)\right) d\xi dy_1 \dots dy_{3N-1}}{\int \exp\left(-\beta U_i(\xi, y_1, \dots, y_{3N-1})\right) d\xi dy_1 \dots dy_{3N-1}}, \\
&= \left\langle \exp\left(-\beta V_i(\xi)\right) \right\rangle,
\end{aligned} \tag{4.6}$$

where the angle brackets denote the ensemble average value of $\exp(-\beta V_i(\xi))$ for the system. Thus the unbiased probability $p_i(\xi)$ in (4.5) becomes

$$\begin{aligned}
p_i(\xi) &= p_i^*(\xi) \cdot \exp(\beta V_i(\xi)) \cdot \left\langle \exp(-\beta V_i(\xi)) \right\rangle \\
&\equiv p_i^*(\xi) \cdot \exp(\beta V_i(\xi)) \cdot \exp(\beta F_i),
\end{aligned} \tag{4.7}$$

which defines the quantity F_i , the average free energy introduced by adding the biasing potential $V_i(\xi)$. We have therefore formulated the unbiased probability $p_i(\xi)$ as an explicit function of the biased probability $p_i^*(\xi)$, which we obtain from our umbrella sampling windows. The value of $V_i(\xi)$ can be computed directly from its definition. One is then left with the quantity F_i for each window. These can be computed directly from the data using the Weighted Histogram Analysis Method (214), for which there exist standard implementations (178).

Bibliography

1. Petsko GA & Ringe D (2004) *Protein structure and function* (New Science Press, London).
2. Richardson JS (1981) The anatomy and taxonomy of protein structure. *Adv Protein Chem* 34:167-339.
3. Wang Z & Moult J (2001) SNPs, protein structure, and disease. *Human Mutation* 17(4):263-270.
4. Vihinen M (1987) Relationship of protein flexibility to thermostability. *Protein Engineering* 1(6):477-480.
5. Falke JJ (2002) A Moving Story. *Science* 295(5559):1480-1481.
6. Fisher CK & Stultz CM (2011) Protein Structure along the Order–Disorder Continuum. *Journal of the American Chemical Society* 133(26):10022-10025.
7. Chouard T (2011) Structural biology: Breaking the protein rules. in *Nature*, pp 151-153.
8. Conway KA, *et al.* (2000) Acceleration of oligomerization, not fibrillization, is a shared property of both alpha-synuclein mutations linked to early-onset Parkinson's disease: Implications for pathogenesis and therapy. *Proceedings of the National Academy of Sciences* 97(2):571-576.
9. Perutz MF (1999) Glutamine repeats and neurodegenerative diseases: molecular aspects. *Trends in Biochemical Sciences* 24(2):58-63.
10. Romero P, *et al.* (2001) Sequence complexity of disordered protein. *Proteins: Structure, Function, and Bioinformatics* 42(1):38-48.
11. Crick S, Jayaraman M, Frieden C, Wetzel R, & Pappu R (2006) Fluorescence correlation spectroscopy shows that monomeric polyglutamine molecules form collapsed structures in aqueous solutions. *Proc Natl Acad Sci USA* 103(45):16764 - 16769.
12. Vitalis A, Lyle N, & Pappu RV (2009) Thermodynamics of β -Sheet Formation in Polyglutamine. *Biophysical Journal* 97(1):303-311.
13. Hardy J & Selkoe DJ (2002) The Amyloid Hypothesis of Alzheimer's Disease: Progress and Problems on the Road to Therapeutics. *Science* 297(5580):353-356.
14. Pitschke M, Prior R, Haupt M, & Riesner D (1998) Detection of single amyloid beta-protein aggregates in the cerebrospinal fluid of Alzheimer's patients by fluorescence correlation spectroscopy. *Nat Med* 4(7):832-834.
15. El-Agnaf OMA, Mahil DS, Patel BP, & Austen BM (2000) Oligomerization and Toxicity of β -Amyloid-42 Implicated in Alzheimer's Disease. *Biochemical and Biophysical Research Communications* 273(3):1003-1007.
16. Zhang S, *et al.* (2000) The Alzheimer's Peptide A β Adopts a Collapsed Coil Structure in Water. *Journal of Structural Biology* 130(2–3):130-141.
17. Schweers O, Schönbrunn-Hanebeck E, Marx A, & Mandelkow E (1994) Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure. *Journal of Biological Chemistry* 269(39):24290-24297.
18. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, & Dunker AK (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* 323(3):573-584.
19. Joerger AC & Fersht AR (2010) The tumor suppressor p53: from structures to drug discovery. in *Cold Spring Harb Perspect Biol* (Cold Spring Harbor Lab), pp a000919-a000919.
20. Campen A, *et al.* (2008) TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. in *Protein Pept. Lett.*, pp 956-963.
21. Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181(4096):223-230.

22. Huang A & Stultz CM (2009) Finding order within disorder: elucidating the structure of proteins associated with neurodegenerative disease. *Future Med Chem* 1(3):467-482.
23. Leopold PE, Montal M, & Onuchic JN (1992) Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc Natl Acad Sci U S A* 89(18):8721-8725.
24. Onuchic JN & Wolynes PG (2004) Theory of protein folding. *Curr Opin Struct Biol* 14(1):70-75.
25. Chen J (2012) Towards the physical basis of how intrinsic disorder mediates protein function. in *Archives of Biochemistry and Biophysics*, pp 123-131.
26. Tompa P (2011) Unstructural biology coming of age. *Curr Opin Struct Biol.* 3:419-425.
27. Dunker A, Cortese M, Romero P, Iakoucheva L, & Uversky V (2005) Flexible nets. The roles of intrinsic disorder in protein interaction networks. *The FEBS journal* 272(20):5129 - 5148.
28. Uversky VN (2002) Natively unfolded proteins: A point where biology waits for physics. *Protein Science* 11(4):739-756.
29. Oldfield C, *et al.* (2008) Flexible nets: Disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics* 9(S1):S1.
30. Wright PE & Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 293(2):321-331.
31. Gross M (2011) Anarchy in the proteome. *Chemistry World* 8(8):42-45.
32. Monod J, Wyman J, & Changeux J-P (1965) On the nature of allosteric transitions: A plausible model. in *Journal of Molecular Biology*, pp 88-118.
33. Kumar S, Showalter SA, & Noid WG (2013) Native-Based Simulations of the Binding Interaction Between RAP74 and the Disordered FCP1 Peptide. *The Journal of Physical Chemistry B*:1-12.
34. Shoemaker BA, Portman JJ, & Wolynes PG (2000) Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc Natl Acad Sci U S A* 97(16):8868-8873.
35. Levy Y, Onuchic JN, & Wolynes PG (2007) Fly-casting in protein-DNA binding: frustration between protein folding and electrostatics facilitates target recognition. *J Am Chem Soc* 129(4):738-739.
36. MacArthur MW, Driscoll PC, & Thornton JM (1994) NMR and crystallography--complementary approaches to structure determination. *Trends Biotechnol* 12(5):149-153.
37. Kjaergaard M, Teilum K, & Poulsen FM (2010) Conformational selection in the molten globule state of the nuclear coactivator binding domain of CBP. *Proceedings of the National Academy of Sciences of the United States of America* 107(28):12535-12540.
38. Kosol S, Contreras-Martos S, Cedeno C, & Tompa P (2013) Structural characterization of intrinsically disordered proteins by NMR spectroscopy. *Molecules* 18(9):10802-10828.
39. Gillespie JR & Shortle D (1997) Characterization of long-range structure in the denatured state of staphylococcal nuclease. I. Paramagnetic relaxation enhancement by nitroxide spin labels. *J Mol Biol* 268(1):158-169.
40. Dedmon MM, Lindorff-Larsen K, Christodoulou J, Vendruscolo M, & Dobson CM (2005) Mapping long-range interactions in alpha-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *J Am Chem Soc* 127(2):476-477.
41. Eliezer D (2009) Biophysical characterization of intrinsically disordered proteins. *Curr Opin Struct Biol* 19(1):23-30.
42. Bernado P & Svergun DI (2012) Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol Biosyst* 8(1):151-167.
43. Ando T (2013) High-speed atomic force microscopy. *Microscopy (Oxf)* 62(1):81-93.
44. Katan AJ & Dekker C (2011) High-speed AFM reveals the dynamics of single biomolecules at the nanometer scale. *Cell* 147(5):979-982.
45. Miyagi A, *et al.* (2008) Visualization of intrinsically disordered regions of proteins by high-speed atomic force microscopy. *Chemphyschem* 9(13):1859-1866.

46. Ando T, Uchihashi T, & Scheuring S (2014) Filming biomolecular processes by high-speed atomic force microscopy. *Chem Rev* 114(6):3120-3188.
47. Rauscher S & Pomès R (2010) Molecular simulations of protein disorder. *Biochem. Cell Biol.* 88(2):269-290.
48. Tuckerman ME (2010) *Statistical mechanics : theory and molecular simulation* (Oxford University Press, Oxford ; New York) pp xv, 696 p.
49. Karplus M & McCammon JA (2002) Molecular dynamics simulations of biomolecules. in *Nature Structural Biology*.
50. Leach AR (2001) *Molecular modelling: principles and applications* (Prentice Hall, Harlow, England; New York) 2nd Ed pp xxiv, 744 p., 716 p. of plates.
51. Vitalis A & Pappu RV (2009) ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *Journal of Computational Chemistry* 30(5):673-699.
52. Vitalis A & Caflisch A (2010) Micelle-like architecture of the monomer ensemble of Alzheimer's amyloid-beta peptide in aqueous solution and its implications for Abeta aggregation. *J Mol Biol* 403(1):148-165.
53. Meng W, Lyle N, Luan B, Raleigh DP, & Pappu RV (2013) Experiments and simulations show how long-range contacts can form in expanded unfolded proteins with negligible secondary structure. *Proc Natl Acad Sci U S A* 110(6):2123-2128.
54. Wuttke R, et al. (2014) Temperature-dependent solvation modulates the dimensions of disordered proteins. *Proc Natl Acad Sci U S A* 111(14):5213-5218.
55. Bottaro S, Lindorff-Larsen K, & Best RB (2013) Variational Optimization of an All-Atom Implicit Solvent Force Field to Match Explicit Solvent Simulation Data. *J Chem Theory Comput* 9(12):5641-5652.
56. Jha AK, Colubri A, Freed KF, & Sosnick TR (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.
57. Ozenne V, et al. (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.
58. Bernstein FC, et al. (1977) The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology* 112(3):535-542.
59. Vendruscolo M (2007) Determination of conformationally heterogeneous states of proteins. *Current Opinion in Structural Biology* 17(1):15-20.
60. Dedmon MM, Lindorff-Larsen K, Christodoulou J, Vendruscolo M, & Dobson CM (2004) 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.
61. Fisher CK & Stultz CM (2011) Constructing ensembles for intrinsically disordered proteins. *Current Opinion in Structural Biology* 21(3):426-431.
62. Varadi M, et al. (2014) pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Research* 42(D1):D326-D335.
63. Krzeminski M, Marsh JA, Neale C, Choy W-Y, & Forman-Kay JD (2013) Characterization of disordered proteins with ENSEMBLE. *Bioinformatics* 29(3):398-399.
64. Jensen MR, Salmon L, Nodet G, & Blackledge M (2010) Defining Conformational Ensembles of Intrinsically Disordered and Partially Folded Proteins Directly from Chemical Shifts. *Journal of the American Chemical Society* 132(4):1270-1272.
65. Fisher CK, Huang A, & Stultz CM (2010) Modeling Intrinsically Disordered Proteins with Bayesian Statistics. *Journal of the American Chemical Society* 132(42):14919-14927.

66. Marsh JA & Forman-Kay JD (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.
67. Ganguly D & Chen JH (2009) Structural Interpretation of Paramagnetic Relaxation Enhancement-Derived Distances for Disordered Protein States. *Journal of Molecular Biology* 390(3):467-477.
68. Huang A & Stultz CM (2008) The Effect of a Delta K280 Mutation on the Unfolded State of a Microtubule-Binding Repeat in Tau. *Plos Computational Biology* 4(8).
69. Pitera JW & Chodera JD (2012) On the Use of Experimental Observations to Bias Simulated Ensembles. *Journal of Chemical Theory and Computation* 8(10):3445-3451.
70. Boomsma W, Ferkinghoff-Borg J, & Lindorff-Larsen K (2014) Combining experiments and simulations using the maximum entropy principle. *PLoS Comput Biol* 10(2):e1003406.
71. Lane JL, Schwantes CR, A. BK, & Pande VS (2014) Efficient inference of protein structural ensembles. *arXiv:1408.0255v1*.
72. Fisher CK, Ullman, Orly, and Stultz, Collin M. (2012) Efficient Construction of Disordered Protein Ensembles in a Bayesian Framework with Optimal Selection of Conformations. *Pacific Symposium on Biocomputing* 17:82-93.
73. Neal S, Nip AM, Zhang HY, & Wishart DS (2003) Rapid and accurate calculation of protein H-1, C-13 and N-15 chemical shifts. *J Biomol NMR* 26(3):215-240.
74. Gurry T, et al. (2013) The Dynamic Structure of alpha-Synuclein Multimers. *Journal of the American Chemical Society*.
75. Ross CA & Poirier MA (2004) Protein aggregation and neurodegenerative disease. *Nature Medicine* 10:S10-S17.
76. Cohen SIA, et al. (2013) Proliferation of amyloid-β42 aggregates occurs through a secondary nucleation mechanism. *Proceedings of the National Academy of Sciences* 110(24):9758-9763.
77. Kaye R, et al. (2003) Common Structure of Soluble Amyloid Oligomers Implies Common Mechanism of Pathogenesis. *Science* 300(5618):486-489.
78. Walsh DM & Selkoe DJ (2007) Abeta Oligomers – a decade of discovery. *Journal of Neurochemistry* 101(5):1172-1184.
79. Periquet M, Fulga T, Myllykangas L, Schlossmacher MG, & Feany MB (2007) Aggregated alpha-Synuclein Mediates Dopaminergic Neurotoxicity In Vivo. *The Journal of Neuroscience* 27(12):3338-3346.
80. Bernstein SL, et al. (2009) Amyloid-beta protein oligomerization and the importance of tetramers and dodecamers in the aetiology of Alzheimer's disease. *Nat Chem* 1(4):326-331.
81. Laganowsky A, et al. (2012) Atomic View of a Toxic Amyloid Small Oligomer. *Science* 335(6073):1228-1231.
82. Glenner GG & Wong CW (1984) Alzheimer's disease: Initial report of the purification and characterization of a novel cerebrovascular amyloid protein. *Biochemical and Biophysical Research Communications* 120(3):885-890.
83. Haass C & Selkoe DJ (2007) Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid beta-peptide. *Nat Rev Mol Cell Biol* 8(2):101-112.
84. Haass C (2004) Take five—BACE and the gamma-secretase quartet conduct Alzheimer's amyloid beta-peptide generation. *The EMBO Journal* 23(3):483-488.
85. Burdick D, et al. (1992) Assembly and aggregation properties of synthetic Alzheimer's Abeta amyloid peptide analogs. *Journal of Biological Chemistry* 267(1):546-554.
86. Citron M, et al. (1992) Mutation of the beta-amyloid precursor protein in familial Alzheimer's disease increases beta-protein production. *Nature* 360(6405):672-674.
87. Kassler K, Horn AC, & Sticht H (2010) Effect of pathogenic mutations on the structure and dynamics of Alzheimer's Aβ42-amyloid oligomers. *J Mol Model* 16(5):1011-1020.

88. Suzuki N, *et al.* (1994) An increased percentage of long amyloid beta protein secreted by familial amyloid beta protein precursor (beta APP717) mutants. *Science* 264(5163):1336-1340.
89. Kirkitadze MD, Condrón MM, & Teplow DB (2001) Identification and characterization of key kinetic intermediates in amyloid beta-protein fibrillogenesis. *Journal of Molecular Biology* 312(5):1103-1119.
90. Glabe CG (2006) Common mechanisms of amyloid oligomer pathogenesis in degenerative disease. *Neurobiology of Aging* 27(4):570-575.
91. Selkoe DJ (2003) Folding proteins in fatal ways. *Nature* 426(6968):900-904.
92. Bartels T, CJG, Selkoe D.J. (2011) alpha-Synuclein occurs physiologically as a helically folded tetramer that resists aggregation. *Nature* 477:107-110.
93. Lashuel HA, Hartley D, Petre BM, Walz T, & Lansbury PT (2002) Amyloid pores from pathogenic mutations. *Nature* 418.
94. Demuro A, *et al.* (2005) Calcium Dysregulation and Membrane Disruption as a Ubiquitous Neurotoxic Mechanism of Soluble Amyloid Oligomers. *Journal of Biological Chemistry* 280(17):17294-17300.
95. Silvia C, *et al.* (2010) A causative link between the structure of aberrant protein oligomers and their toxicity. *Nature Chemical Biology* 6(2):140-147.
96. Sandberg A, *et al.* (2010) Stabilization of neurotoxic Alzheimer amyloid- β oligomers by protein engineering. *Proceedings of the National Academy of Sciences* 107(35):15595-15600.
97. Hoyer W, Grönwall C, Jonsson A, Ståhl S, & Härd T (2008) Stabilization of a beta-hairpin in monomeric Alzheimer's amyloid-beta peptide inhibits amyloid formation. *Proceedings of the National Academy of Sciences* 105(13):5099-5104.
98. Ahmed M, *et al.* (2010) Structural conversion of neurotoxic amyloid-beta1-42 oligomers to fibrils. *Nat Struct Mol Biol* 17(5):561-567.
99. Bitan G, *et al.* (2003) Amyloid β -protein (A β) assembly: A β 40 and A β 42 oligomerize through distinct pathways. *Proceedings of the National Academy of Sciences* 100(1):330-335.
100. Bitan G, Vollers SS, & Teplow DB (2003) Elucidation of Primary Structure Elements Controlling Early Amyloid β -Protein Oligomerization. *Journal of Biological Chemistry* 278(37):34882-34889.
101. Bernstein SL, *et al.* (2009) Amyloid- β protein oligomerization and the importance of tetramers and dodecamers in the aetiology of Alzheimer's disease. *Nat Chem* 1(4):326-331.
102. Urbanc B, *et al.* (2004) In silico study of amyloid β -protein folding and oligomerization. *Proceedings of the National Academy of Sciences of the United States of America* 101(50):17345-17350.
103. Urbanc B, Betnel M, Cruz L, Bitan G, & Teplow DB (2010) Elucidation of Amyloid β -Protein Oligomerization Mechanisms: Discrete Molecular Dynamics Study. *Journal of the American Chemical Society* 132(12):4266-4280.
104. Urbanc B, *et al.* (2011) Structural Basis for A β 1-42 Toxicity Inhibition by A β C-Terminal Fragments: Discrete Molecular Dynamics Study. *Journal of Molecular Biology* 410(2):316-328.
105. Spillantini M.G., SML, Lee V.M.-Y., Trojanowski J.Q., Jakes R., Goedert M. (1997) Alpha-synuclein in Lewy bodies. *Nature* 388:839-840.
106. Karran E, Mercken M, & Strooper BD (2011) The amyloid cascade hypothesis for Alzheimer's disease: an appraisal for the development of therapeutics. *Nat Rev Drug Discov* 10(9):698-712.
107. Chiti F & Dobson CM (2006) Protein Misfolding, Functional Amyloid, and Human Disease. *Annual Review of Biochemistry* 75(1):333-366.
108. Petkova AT, Yau W-M, & Tycko R (2005) Experimental Constraints on Quaternary Structure in Alzheimer's beta-Amyloid Fibrils. *Biochemistry* 45(2):498-512.

109. Paravastu AK, Leapman RD, Yau W-M, & Tycko R (2008) Molecular structural basis for polymorphism in Alzheimer's beta-amyloid fibrils. *Proceedings of the National Academy of Sciences* 105(47):18349-18354.
110. Lührs T, et al. (2005) 3D structure of Alzheimer's amyloid- β (1–42) fibrils. *Proceedings of the National Academy of Sciences of the United States of America* 102(48):17342-17347.
111. Lu J-X, et al. (2013) Molecular Structure of Beta-Amyloid Fibrils in Alzheimer's Disease Brain Tissue. *Cell* 154(6):1257-1268.
112. Rubin N, Perugia E, Goldschmidt M, Fridkin M, & Addadi L (2008) Chirality of Amyloid Suprastructures. *Journal of the American Chemical Society* 130(14):4602-4603.
113. GhattyVenkataKrishna PK, Uberbacher EC, & Cheng X (2013) Effect of the amyloid beta hairpin's structure on the handedness of helices formed by its aggregates. *FEBS Letters* 587(16):2649-2655.
114. Buchete N-V, Tycko R, & Hummer G (2005) Molecular Dynamics Simulations of Alzheimer's β -Amyloid Protofilaments. *Journal of Molecular Biology* 353(4):804-821.
115. Fawzi NL, Okabe Y, Yap E-H, & Head-Gordon T (2007) Determining the Critical Nucleus and Mechanism of Fibril Elongation of the Alzheimer's A β 1–40 Peptide. *Journal of Molecular Biology* 365(2):535-550.
116. Lapidus LJ (2013) Exploring the top of the protein folding funnel by experiment. *Current Opinion in Structural Biology* 23(1):30-35.
117. Ahmad B, Chen Y, & Lapidus LJ (2012) Aggregation of alpha-synuclein is kinetically controlled by intramolecular diffusion. *Proceedings of the National Academy of Sciences* 109(7):2336-2341.
118. Naiki H & Nakakuki K (1996) First-order kinetic model of Alzheimer's beta-amyloid fibril extension in vitro. *Laboratory investigation; a journal of technical methods and pathology* 74(2):374-383.
119. Meyer-Luehmann M, et al. (2006) Exogenous Induction of Cerebral β -Amyloidogenesis Is Governed by Agent and Host. *Science* 313(5794):1781-1784.
120. Eisele YS, et al. (2010) Peripherally Applied A β -Containing Inoculates Induce Cerebral β -Amyloidosis. *Science* 330(6006):980-982.
121. Fisher Charles K, Ullman O, & Stultz Collin M (2013) Comparative Studies of Disordered Proteins with Similar Sequences: Application to A β 40 and A β 42. *Biophysical Journal* 104(7):1546-1555.
122. Fitzpatrick AWP, et al. (2013) Atomic structure and hierarchical assembly of a cross-beta amyloid fibril. *Proceedings of the National Academy of Sciences* 110(14):5468-5473.
123. Lee J, Culyba EK, Powers ET, & Kelly JW (2011) Amyloid- β forms fibrils by nucleated conformational conversion of oligomers. *Nat Chem Biol* 7(9):602-609.
124. Teplow DB, et al. (2006) Elucidating Amyloid β -Protein Folding and Assembly: A Multidisciplinary Approach. *Accounts of Chemical Research* 39(9):635-645.
125. Lin Y-S, Bowman Gregory R, Beauchamp Kyle A, & Pande Vijay S (2012) Investigating How Peptide Length and a Pathogenic Mutation Modify the Structural Ensemble of Amyloid Beta Monomer. *Biophysical Journal* 102(2):315-324.
126. Bellucci A, et al. (2012) From alpha-synuclein to synaptic dysfunctions: New insights into the pathophysiology of Parkinson's disease. *Brain Research*.
127. Uversky VN, Li J, & Fink AL (2001) Evidence for a Partially Folded Intermediate in alpha-Synuclein Fibril Formation. *Journal of Biological Chemistry* 276(14):10737-10744.
128. Bucciantini M, et al. (2002) Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* 416(6880):507-511.
129. Danzer KM, et al. (2007) Different Species of alpha-Synuclein Oligomers Induce Calcium Influx and Seeding. *The Journal of Neuroscience* 27(34):9220-9232.

130. Winner B, *et al.* (2011) In vivo demonstration that alpha-synuclein oligomers are toxic. *Proceedings of the National Academy of Sciences* 108(10):4194-4199.
131. Uversky V (2003) A protein-chameleon: conformational plasticity of alpha-synuclein, a disordered protein involved in neurodegenerative disorders. *J Biomol Struct Dyn* 21(2):211 - 234.
132. Drescher M, Huber M, & Subramaniam V (2012) Hunting the Chameleon: Structural Conformations of the Intrinsically Disordered Protein Alpha-Synuclein. *ChemBioChem* 13(6):761-768.
133. Fauvet B, *et al.* (2012) Alpha-synuclein in the central nervous system and from erythrocytes, mammalian cells and E. coli exists predominantly as a disordered monomer. *Journal of Biological Chemistry*.
134. Wang W, *et al.* (2011) A soluble alpha-synuclein construct forms a dynamic tetramer. *Proceedings of the National Academy of Sciences* 108(43):17797-17802.
135. Weinreb P, Zhen W, Poon A, Conway K, & Lansbury P (1996) NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. *Biochemistry* 35(43):13709 - 13715.
136. Trexler AJ & Rhoades E (2012) N-terminal acetylation is critical for forming α -helical oligomer of α -synuclein. *Protein Science* 21(5):601-605.
137. Binolfi A, Theillet F, & Selenko P (2012) Bacterial in-cell NMR of human alpha-synuclein: a disordered monomer by nature? *Biochemical Society Transactions* 40:950-954.
138. Ullman O, Fisher CK, & Stultz CM (2011) Explaining the Structural Plasticity of alpha-Synuclein. *Journal of the American Chemical Society* 133(48):19536-19546.
139. Mukherjea M, *et al.* (2009) Myosin VI Dimerization Triggers an Unfolding of a Three-Helix Bundle in Order to Extend Its Reach. *Molecular Cell* 35(3):305-315.
140. Lawson DM, *et al.* (1991) Solving the structure of human H ferritin by genetically engineering intermolecular crystal contacts. *Nature* 349(6309):541-544.
141. Berman H, *et al.* (1999) The Protein Data Bank. *Nucleic Acids Research* 28(1):235-242.
142. Sugita Y & Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* 314:141-151.
143. Lazaridis T & Karplus M (1999) Effective energy function for proteins in solution. *Proteins: Structure, Function, and Bioinformatics* 35(2):133-152.
144. Brooks BR, *et al.* (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry* 4(2):187-217.
145. Neal S, Nip A, Zhang H, & Wishart D (2003) Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *J Biomol NMR* 26(3):215-240.
146. Zweckstetter M (2008) NMR: prediction of molecular alignment from structure using the PALES software. *Nat. Protocols* 3(4):679-690.
147. Kabsch W & Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22(12):2577-2637.
148. Kay L, Keifer P, & Saarinen T (1992) Pure absorption gradient enhanced heteronuclear single quantum correlation spectroscopy with improved sensitivity. *Journal of the American Chemical Society* 114(26):10663-10665.
149. Goddard TD & Kneller DG (SPARKY 3 (University of California, San Francisco).
150. Bax A (2003) Weak alignment offers new NMR opportunities to study protein structure and dynamics. *Protein Science* 12(1):1-16.
151. Schwieters CD, Kuszewski JJ, Tjandra N, & Marius Clore G (2003) The Xplor-NIH NMR molecular structure determination package. *Journal of Magnetic Resonance* 160(1):65-73.
152. Schrodinger, LLC (2010) The PyMOL Molecular Graphics System, Version 1.3r1.

153. Wishart D, Bigam C, Holm A, Hodges R, & Sykes B (1995) ^1H , ^{13}C and ^{15}N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. *J Biomol NMR* 5(1):67-81.
154. el-Agnaf OM & Irvine GB (2002) Aggregation and neurotoxicity of alpha-synuclein and related peptides. *Biochemical Society transactions* 30(4):559-565.
155. Volles MJ, *et al.* (2001) Vesicle Permeabilization by Protofibrillar alpha-Synuclein: Implications for the Pathogenesis and Treatment of Parkinson's Disease. *Biochemistry* 40(26):7812-7819.
156. Cho M-K, *et al.* (2009) Structural characterization of α -synuclein in an aggregation prone state. *Protein Science* 18(9):1840-1846.
157. Wu K-P, Kim S, Fela DA, & Baum J (2008) Characterization of Conformational and Dynamic Properties of Natively Unfolded Human and Mouse alpha-Synuclein Ensembles by NMR: Implication for Aggregation. *Journal of Molecular Biology* 378(5):1104-1115.
158. Bartels T, *et al.* (2010) The N-Terminus of the Intrinsically Disordered Protein α -Synuclein Triggers Membrane Binding and Helix Folding. *Biophysical Journal* 99(7):2116-2124.
159. Bodner CR, Dobson CM, & Bax A (2009) Multiple Tight Phospholipid-Binding Modes of α -Synuclein Revealed by Solution NMR Spectroscopy. *Journal of Molecular Biology* 390(4):775-790.
160. Vamvaca K, Volles MJ, & Lansbury Jr PT (2009) The First N-terminal Amino Acids of α -Synuclein Are Essential for α -Helical Structure Formation In Vitro and Membrane Binding in Yeast. *Journal of Molecular Biology* 389(2):413-424.
161. Kang L, *et al.* (2012) N-terminal acetylation of alpha-synuclein induces increased transient helical propensity and decreased aggregation rates in the intrinsically disordered monomers. *Protein Science* 21(7):911-917.
162. Manavalan P & Johnson WC (1985) Protein secondary structure from circular dichroism spectra. *J Biosci* 8(1-2):141-149.
163. Greenfield NJ (2007) Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. *Nat. Protocols* 1(6):2527-2535.
164. Iwai A, *et al.* (1995) The precursor protein of non-A β component of Alzheimer's disease amyloid is a presynaptic protein of the central nervous system. *Neuron* 14(2):467-475.
165. Jarrett JT, Berger EP, & Lansbury PT (1993) The carboxy terminus of the β amyloid protein is critical for the seeding of amyloid formation: Implications for the pathogenesis of Alzheimer's disease. *Biochemistry* 32(18):4693-4697.
166. Yan Y & Wang C (2006) A β 42 is More Rigid than A β 40 at the C Terminus: Implications for A β Aggregation and Toxicity. *Journal of Molecular Biology* 364(5):853-862.
167. Roher AE, *et al.* (1993) beta-Amyloid-(1-42) is a major component of cerebrovascular amyloid deposits: implications for the pathology of Alzheimer disease. *Proceedings of the National Academy of Sciences* 90(22):10836-10840.
168. Jarrett JT, Berger EP, & Lansbury PT (1993) The C-Terminus of the β Protein is Critical in Amyloidogenesis. *Annals of the New York Academy of Sciences* 695(1):144-148.
169. Walsh DM, *et al.* (2002) Naturally secreted oligomers of amyloid beta protein potently inhibit hippocampal long-term potentiation in vivo. *Nature* 416(6880):535-539.
170. Schnabel J (2011) Amyloid: Little proteins, big clues. *Nature* 475(7355):S12-S14.
171. Brooks BR, *et al.* (2009) CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry* 30(10):1545-1614.
172. Dashti DS & Roitberg AE (2013) Optimization of Umbrella Sampling Replica Exchange Molecular Dynamics by Replica Positioning. *Journal of Chemical Theory and Computation* 9(11):4692-4699.
173. Wang DQ, Amundadottir ML, van Gunsteren WF, & Hunenberger PH (2013) Intramolecular hydrogen-bonding in aqueous carbohydrates as a cause or consequence of conformational

- preferences: a molecular dynamics study of cellobiose stereoisomers. *Eur. Biophys. J. Biophys. Lett.* 42(7):521-537.
174. Wilhelm M, *et al.* (2012) Multistep Drug Intercalation: Molecular Dynamics and Free Energy Studies of the Binding of Daunomycin to DNA. *Journal of the American Chemical Society* 134(20):8588-8596.
 175. Huang A & Stultz CM (2007) Conformational sampling with implicit solvent models: Application to the PHF6 peptide in tau protein. *Biophysical Journal* 92(1):34-45.
 176. Strodel B & Wales DJ (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.
 177. Evans DJ & Holian BL (1985) The Nose-Hoover thermostat. *The Journal of Chemical Physics* 83(8):4069-4074.
 178. Grossfield A (2013) WHAM: the weighted histogram analysis method <http://membrane.urmc.rochester.edu/content/wham>, 2.0.7.
 179. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, & Kollman PA (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry* 13(8):1011-1021.
 180. Roux B (1995) The calculation of the potential of mean force using computer simulations. *Computer Physics Communications* 91(1-3):275-282.
 181. Sciarretta KL, Gordon DJ, Petkova AT, Tycko R, & Meredith SC (2005) A β 40-Lactam(D23/K28) Models a Conformation Highly Favorable for Nucleation of Amyloid \dagger . *Biochemistry* 44(16):6003-6014.
 182. Reddy G, Straub JE, & Thirumalai D (2009) Influence of Preformed Asp23-Lys28 Salt Bridge on the Conformational Fluctuations of Monomers and Dimers of A β Peptides with Implications for Rates of Fibril Formation. *The Journal of Physical Chemistry B* 113(4):1162-1172.
 183. Tarus B, Straub JE, & Thirumalai D (2006) Dynamics of Asp23-Lys28 Salt-Bridge Formation in A β 10-35 Monomers. *Journal of the American Chemical Society* 128(50):16159-16168.
 184. Nguyen PH, Li MS, Stock G, Straub JE, & Thirumalai D (2007) Monomer adds to preformed structured oligomers of A beta-peptides by a two-stage dock-lock mechanism. *Proceedings of the National Academy of Sciences of the United States of America* 104(1):111-116.
 185. Takeda T & Klimov DK (2009) Probing Energetics of Abeta Fibril Elongation by Molecular Dynamics Simulations. *Biophysical Journal* 96(11):4428-4437.
 186. Rojas A, Liwo A, Browne D, & Scheraga HA (2010) Mechanism of Fiber Assembly: Treatment of Abeta Peptide Aggregation with a Coarse-Grained United-Residue Force Field. *Journal of Molecular Biology* 404(3):537-552.
 187. Esler WP, *et al.* (2000) Alzheimer's disease amyloid propagation by a template-dependent dock-lock mechanism. *Biochemistry* 39(21):6288-6295.
 188. Elenewski JE & Hackett JC (2010) Free Energy Landscape of the Retinol/Serum Retinol Binding Protein Complex: A Biological Host-Guest System. *Journal of Physical Chemistry B* 114(34):11315-11322.
 189. Zhang DQ, Gullingsrud J, & McCammon JA (2006) Potentials of mean force for acetylcholine unbinding from the alpha7 nicotinic acetylcholine receptor ligand-binding domain. *Journal of the American Chemical Society* 128(9):3019-3026.
 190. Rashid MH & Kuyucak S (2014) Free Energy Simulations of Binding of HsTx1 Toxin to Kv1 Potassium Channels: the Basis of Kv1.3/Kv1.1 Selectivity. *Journal of Physical Chemistry B* 118(3):707-716.
 191. Dinner AR & Karplus M (1999) Is protein unfolding the reverse of protein folding? A lattice simulation analysis. *Journal of Molecular Biology* 292(2):403-419.

192. Daggett V (2002) Molecular Dynamics Simulations of the Protein Unfolding/Folding Reaction. *Accounts of Chemical Research* 35(6):422-429.
193. Fawzi NL, Ying J, Ghirlando R, Torchia DA, & Clore GM (2011) Atomic-resolution dynamics on the surface of amyloid-beta protofibrils probed by solution NMR. *Nature* 480(7376):268-272.
194. Kumar-Singh S, et al. (2002) In Vitro Studies of Flemish, Dutch, and Wild-Type β -Amyloid Provide Evidence for Two-Staged Neurotoxicity. *Neurobiology of Disease* 11(2):330-340.
195. Nilsberth C, et al. (2001) The 'Arctic' APP mutation (E693G) causes Alzheimer's disease by enhanced A β protofibril formation. *Nature neuroscience* 4(9):887-893.
196. Langer F, et al. (2011) Soluble A β seeds are potent inducers of cerebral beta-amyloid deposition. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 31(41):14488-14495.
197. Luca S, Yau W-M, Leapman R, & Tycko R (2007) Peptide Conformation and Supramolecular Organization in Amylin Fibrils: Constraints from Solid-State NMR†. *Biochemistry* 46(47):13505-13522.
198. Cooper GJ, et al. (1987) Purification and characterization of a peptide from amyloid-rich pancreases of type 2 diabetic patients. *Proceedings of the National Academy of Sciences* 84(23):8628-8632.
199. Westermark P, et al. (1987) Amyloid fibrils in human insulinoma and islets of Langerhans of the diabetic cat are derived from a neuropeptide-like protein also present in normal islet cells. *Proceedings of the National Academy of Sciences* 84(11):3881-3885.
200. Casalone C, et al. (2004) Identification of a second bovine amyloidotic spongiform encephalopathy: Molecular similarities with sporadic Creutzfeldt-Jakob disease. *Proceedings of the National Academy of Sciences of the United States of America* 101(9):3065-3070.
201. Yamaguchi H, Hirai S, Morimatsu M, Shoji M, & Harigaya Y (1988) Diffuse type of senile plaques in the brains of Alzheimer-type dementia. *Acta Neuropathol* 77(2):113-119.
202. Goedert M, Spillantini MG, Jakes R, Rutherford D, & Crowther RA (1989) Multiple isoforms of human microtubule-associated protein tau: sequences and localization in neurofibrillary tangles of Alzheimer's disease. *Neuron* 3(4):519-526.
203. Schweers O, Schonbrunn-Hanebeck E, Marx A, & Mandelkow E (1994) Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure. *J Biol Chem* 269(39):24290 - 24297.
204. Wetzel R (2006) Kinetics and Thermodynamics of Amyloid Fibril Assembly. *Accounts of Chemical Research* 39(9):671-679.
205. Shewmaker F, Wickner RB, & Tycko R (2006) Amyloid of the prion domain of Sup35p has an in-register parallel beta-sheet structure. *Proceedings of the National Academy of Sciences* 103(52):19754-19759.
206. Bodles AM, Guthrie DJS, Greer B, & Irvine GB (2001) Identification of the region of non-A β component (NAC) of Alzheimer's disease amyloid responsible for its aggregation and toxicity. *Journal of Neurochemistry* 78(2):384-395.
207. Shaw DE, et al. (2008) Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* 51(7):91-97.
208. Starck C & Sutherland-Smith A (2010) Cytotoxic Aggregation and Amyloid Formation by the Myostatin Precursor Protein. *PLoS ONE* 5(2):e9170.
209. Hauser CAE, et al. (2011) Natural tri- to hexapeptides self-assemble in water to amyloid β -type fiber aggregates by unexpected α -helical intermediate structures. *Proceedings of the National Academy of Sciences* 108(4):1361-1366.
210. Auer S, Dobson CM, & Vendruscolo M (2007) Characterization of the nucleation barriers for protein aggregation and amyloid formation. *HFSP Journal* 1(2):137-146.

211. MacKerell A, Wiórkiewicz-Kuczera J, & Karplus M (1995) CHARMM22 Parameter Set. *Harvard University Department of Chemistry, Cambridge, MA.*
212. Chen B-C, *et al.* (2014) Lattice light-sheet microscopy: Imaging molecules to embryos at high spatiotemporal resolution. *Science* 346(6208).
213. Blondel A & Karplus M (1996) New formulation for derivatives of torsion angles and improper torsion angles in molecular mechanics: Elimination of singularities. *Journal of Computational Chemistry* 17(9):1132-1141.
214. Souaille M & Roux Bt (2001) Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Computer Physics Communications* 135(1):40-57.