Summer 8-15-2013

Quantification of Conformational Heterogeneity and its Role in Protein Aggregation and Unfolding

Nicholas J. Lyle
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences

Computational and Systems Biology

Dissertation Examination Committee:
Rohit V. Pappu, Chair
Jan Bieschke
Anders E. Carlsson
James J. Havranek
Baranidharan Raman
Gary D. Stormo

Quantification of Conformational Heterogeneity and its Role in Protein Aggregation and Unfolding

by

Nicholas J. Lyle

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2013

St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost I would like to thank my thesis mentor Rohit. The past years have been an incredible journey of scientific and personal growth and such experiences come but once in a lifetime but remain with us till the end. I have yet to meet any individual who was more invested in the pursuit of knowledge and I learned from him not only how to answer questions, but what are the right questions to ask. To all of his pursuits he offered a sustained effort set to the maximum of his abilities. This extended to making sure each and every individual of the lab was in the best possible position to maximize their strengths while discovering and being honest about their weaknesses.

I would also like to acknowledge my cohorts as each has shaped my development as a scientist and given me unique insights into life and myself. I owe a huge debt of gratitude to Andreas. He instilled a sense of fearlessness that is necessary to discovery but balanced this with the practical need to produce meaningful results. He also challenged me intellectually and pushed me to be a better thinker and person. Other past members include: Alan who delivered incredible insights, advice, and showed me the joy of just thinking and discussing – about anything; Albert for his honesty and clarity of thought and idealistic approach to things that kept me optimistic about the future; Tim for showing me the value of hard work, self-reflection, and insight into what things are important and what are illusions; Adam for his ability to transform a difficult situation into a lighthearted one; and Matt for interesting discussions. Current members include: Scott, a close ally and confidant that has shown me what it means to be critical of results and provided extensive input that has helped steer the direction of my research; Rahul, another close ally, sounding-board, and voice of reason willing to provide input and

thought on any topic; Kiersten, a hard worker willing to take on any challenge; Anu who is the true example of what it means to be a good and helpful person; and Tyler who keeps us all honest. Several undergraduates have been of assistance, particularly Nil who helped me wrap up and analyze some results included in this document.

Funding came jointly from the NIH (5R01-NS056114) and NSF (MCB 0718924) and I am grateful to these agencies. I would like to thank my thesis committee members for helpful suggestions, insights, and directions that have been necessary to make it to the finish line.

To conclude, I would like to thank my wife Kat for her amazing support. Through the years she has put in monumental effort to help me succeed and to be happy in life. The long road to a doctorate is not an easy one but I am lucky to have her as a companion and true friend by my side to smooth out the path.

ABSTRACT OF THE DISSERTATION

Quantification of Conformational Heterogeneity and its Role

in Protein Aggregation and Unfolding

by

Nicholas J. Lyle

Doctor of Philosophy in Computational and Systems Biology

Washington University in St. Louis, 2013

Professor Rohit V. Pappu, Chair

Proteins can exhibit significant conformational heterogeneity either under denaturing conditions or in aqueous solutions. The latter is true for a class of proteins whose sequences predispose them to form heterogeneous ensembles of conformations. Characterization of conformational heterogeneity in a protein ensemble requires the quantification of the amplitudes of spontaneous fluctuations in conjunction with information regarding coarse grain measures that report on the average sizes, shapes, and densities. This often demands multiplexed experimental approaches whose readouts are interpreted or annotated using ensembles drawn from atomistic or coarse grain computational simulations. Efforts to characterize conformational heterogeneity contribute directly to our understanding of disorder-to-order transitions in protein folding and self-assembly. These efforts are also crucial to our understanding of the heterotypic interactions involving intrinsically disordered proteins and non-native states of well-folded proteins. These heterotypic interactions are important in signal transduction and the regulation of protein homeostasis. The onset and progression of several systemic and

neurodegenerative “conformational diseases” are linked to the nature and degree of conformational heterogeneity in specific proteins or proteolytic products of proteins.

This thesis work focuses on the quantitative characterization of conformational heterogeneity in simulated ensembles of inducibly unfolded and intrinsically disordered proteins. Advances in nuclear magnetic resonance spectroscopy afford the possibility of detailed measurements of inter-residue distances and modulations to the relaxation dynamics of paramagnetic spins that are inserted as probes into a protein. These state-of-the-art measurements show interesting features within denatured state ensembles that cannot be explained using canonical random coil models. Here, we use computer simulations to generate plausible facsimiles of denatured state ensembles that reproduce experimental data and demonstrate that the ensembles that are consistent with the data are characterized by the presence of low-likelihood, long-range intra-chain contacts between hydrophobic groups. When placed in the context of sequence conservation information, it appears that these contacts act as gatekeepers that protect proteins from the deleterious consequences of protein aggregation by sequestering hydrophobic groups in an assortment of intra-chain long-range contacts. We also characterize the nature and degree of conformational heterogeneity in glutamine- and asparagine-rich containing systems. These efforts lead to insights regarding the role of conformational heterogeneity in mediating intermolecular associations that are implicated in aggregation and self-assembly of these systems. Analysis of results from atomistic simulations leads to a phenomenological model for the modulation of conformational heterogeneity and degeneracies of intermolecular interactions by naturally occurring sequences that flank polyglutamine domains.

Finally, we develop a formal order parameter to quantify the conformational heterogeneity in simulated ensembles of proteins. When combined with measures of density and fluctuations thereof, it can be used to provide a complete description of the degree and nature of conformational heterogeneity in different ensembles, thus affording the ability to compare different ensembles to each other while also providing a way to categorize conformational transitions.

Chapter 1

The need for quantitative characterization of protein conformational heterogeneity[1]

1.1 Introduction

Biological science is dominated by the so-called structure-function paradigm. The concept of function being a direct consequence of form follows from every day experience with macroscopic objects. This is amplified in the reductionist approach taken by biologists. Decomposition of a biological system into essential components implies those components fit back together in some ordered and well-defined arrangement. As progress is made in understanding biology at the nano-scales, this structure-centric way of thinking has followed and catalyzed the growth of biophysics. Emil Fischer proposed that enzyme specificity could be explained by shape complementarity [(1, 2)]. He used the metaphor of a lock and key to illustrate how the three-dimensional arrangements of atoms comprising an enzyme and its substrate could enable them to fit together and prevent non-specific catalysis. Interpreted literally, this metaphor suggests that proteins possess rigid structures that determine their functions. Fischer, however, felt that popular interpretations of his lock-and-key metaphor exceeded its scope and experimental justification [(3)]. Indeed, protein rigidity has proven to be unsatisfactory at explaining noncompetitive inhibition and cannot account for enzymes where binding of one

[1] Some parts of the text for this chapter were taken from a recent review co-authored by the candidate. A.H. Mao, N. Lyle, R.V. Pappu. (2013). Describing sequence-ensemble relationships for intrinsically disordered proteins. *Biochemical Journal*, 449: 307-318

reactive group increases the exposure of another [4]. The concepts of allosteric linkage [5] and induced fit require the invocation of protein conformational changes in response to the binding of an interaction partner [6]. The structure-function paradigm, nuanced to accommodate proteins that switch between discrete conformations with different shape complementarities for execution of specific functions provides visual clarity and mathematical simplicity.

1.2 Evidence for function without structure

Advancements on scientific and technological fronts have demonstrated unequivocally that proteins can exhibit significant conformational heterogeneity. Intrinsically disordered proteins (IDPs) are at the extreme end of the heterogeneity spectrum [7-9]. They adopt ensembles of conformations in aqueous solutions for which no single structure or self-similar collection of structures provides an adequate description. By all accounts the conformational heterogeneity exhibited by IDPs is relevant for biological function [8-14]. The phrase intrinsically disordered proteins is used to imply that the amino acid sequences for this class of proteins encode a preference for heterogeneous ensembles of conformations as the thermodynamic ground state under standard physiological conditions (aqueous solutions, 150 mM monovalent salt, low concentrations of divalent ions, pH 7.0, and temperature in the 25ºC – 37ºC range) [9, 15]. The terms *conformational heterogeneity* and *disorder* are sometimes used interchangeably. Thus the classification of a protein as disordered does not imply the protein is entirely devoid of regular secondary structural elements nor does it imply the protein is an expanded random coil.

For many IDPs, folding can be coupled to binding and they can adopt ordered structures in specific bound complexes [16-20]. The intrinsic heterogeneity in their unbound forms is reflected in their ability to adopt different folds in the context of different complexes [10]. Transcription factors represent striking examples of molecules that undergo disorder-to-order

transitions in complex with their cognate DNA partners [21-24]. Highly stable complexes with DNA can make transcription factor dissociation become "unreasonably slow" when compared to the turnover time of downstream regulatory processes. Disorder in the unbound forms is proposed to be important for lowering the overall affinity, which in turn increases the off-rates of protein-DNA complexes [25].

There are a growing number of reports of "fuzzy complexes" whereby conformational heterogeneity prevails in binary and multimolecular complexes [26-28]. IDPs can also self-assemble to form ordered, supramolecular assemblies, although the degree of order within these assemblies is variable and the intermediates that seem to be obligatory for self-assembly are characterized by significant conformational heterogeneity that can be modulated to alter the mechanisms of self-assembly and the stabilities of supramolecular structures [29-38].

1.3 Denatured states of proteins

Although IDPs have gained in interest in recent years, the topic of disorder in denatured proteins has been on the front burner in protein folding studies for over four decades. A folded protein in aqueous milieu can be denatured by addition of cosolutes such as guanidine hydrochloride (GuHCl) or urea; by heating or cooling of the protein plus solvent system; by alterations in the pH of the buffer; or by increased hydrostatic pressure. Denaturing conditions force a protein to unfold by destabilizing the native state or, equivalently, stabilizing the denatured state. Interest in the study of denatured proteins stems from the fact that native state interactions cannot be the only driving force in determining protein stability. Two-state proteins are those that have access to only two thermodynamic macrostates *viz*., the folded and unfolded states. In such systems, one quantifies protein stability as the difference in standard-state chemical potentials associated with the folded and unfolded macrostates. For authentic two-state

proteins, one assumes that the fully unfolded state is accessible via denaturation. Therefore, a typical folding study involves monitoring specific spectroscopic properties as a function of denaturant, which yields a classic folding curve. Assuming congruence between denatured states stabilized in high concentrations of denaturant and the unfolded state, which is seldom populated under physiological conditions, one can estimate the folding free energy of a protein from a folding curve using the so-called linear extrapolation model. Systematic mutational studies provide information regarding the forces that stabilize or destabilize proteins. Interpretation of these results requires knowledge of interactions that are prevalent in folded and unfolded states. The former is accessible through knowledge of the precise three-dimensional structure of a folded protein, while the latter requires computational models for denatured state ensembles.

1.4 Describing conformational heterogeneity

Information encoded at the sequence level keeps IDPs from autonomously folding into singular, well-defined three-dimensional structures [15, 39]. The information content of IDP sequences is such that acquisition of a folded conformation (if this happens) is deferred by coupling the folding process to either binding or self-assembly providing the heterotypic or homotypic interactions in *trans* can stabilize the IDP in a specific fold. From a thermodynamic standpoint, the stabilities of complexes and mechanisms of binding / assembly are linked to the conformational properties of IDPs in their unbound forms. Hence, sequence-ensemble relationships are central to understanding how disorder is used in IDP function. Similarly, quantitative descriptions of conformational heterogeneity in denatured state ensembles (and in non-native states populated under folding conditions for sequences that fold autonomously) are central to our understanding the determinants of protein stability, the description of the effects of

macromolecular crowding and non-specific interactions between a target protein and constituents of cellular milieus, and the active degradation of unfolded proteins within cells.

This thesis work has focused on two classes of protein systems that exhibit significant conformational heterogeneity. These are intrinsically disordered proteins and denatured state ensembles of proteins that do fold as autonomous units. We use a single acronym IDPs to describe both classes of systems where in one case it implies intrinsically disordered proteins and when referring to denatured state ensembles it implies inducible disorder in proteins.

Quantitative descriptions of conformational heterogeneity require biophysical characterization. The signals are often highly averaged and by definition these systems resist crystallization unless they can be forced into specific folded structures. Systems exhibiting conformational heterogeneity also present biochemical challenges because they can be difficult to isolate from tissue or cell systems because the process of homogenization exposes them to proteases that rapidly and preferentially degrade disordered proteins [(40-42)]. Efforts to characterize and quantify conformational heterogeneity have required a systematic integration of biophysical, biochemical, and bioinformatics methods. The major biophysical methodologies include nuclear magnetic resonance (NMR) spectroscopy [(43-50)], steady state and time-resolved fluorescence (single molecule and ensemble) spectroscopies [(51-54)], electron paramagnetic resonance spectroscopy (EPR)[(55)], small angle X-ray scattering (SAXS)[(56-58)], and molecular simulations that are used either *de novo* [(59-67)] or in synergy with data collected from spectroscopic investigations of IDPs [(50, 58, 68-76)].

1.5 The challenge of describing conformational heterogeneity

A single set of position coordinates (and uncertainties in these coordinates) helps relate sequence-to-structure for a protein that folds autonomously into a distinct three-dimensional

structure. Such coordinate sets are generated as models that fit either the electron density data from X-ray diffraction through ordered protein crystals or NMR data that report on the chemical environments of backbone and sidechain protons and nuclei in solution. The protein data bank [(77)] provides a comprehensive archive of coordinate sets for a range of crystallizable proteins and proteins that are amenable for studies by NMR. This rich data set has lead to systematic classification of folds and fold families [(78-81)] thus yielding an improved understanding of sequence-structure relationships and insights regarding the evolution of protein folds. IDPs are not amenable to descriptions by single or even small number distinct of coordinate sets. Instead statistical descriptors are required to provide a concise classification of conformational ensembles and this is the language of polymer physics [(82)]. We focus on these concepts [(82)] to provide a unifying framework for quantitative analysis.

1.6 Limiting models as descriptors of conformational heterogeneity

The two most popular statistical descriptions based on polymer physics are the Flory random coil [(83)] and worm like chain models [(84)]. In the rotational isomeric approximation [(83)] to the Flory random coil model, the conformational partition function for the polypeptide is written as a product of partition functions of independent interaction units. All interactions between non-nearest neighbor units or so-called Kuhn segments are explicitly ignored although the intrinsic conformational preferences of individual units are captured in terms of weights for each of the possible rotational isomers. The unit either spans the degrees of freedom of an individual residue or can take local effects into account to expand the unit to span multiple residues. In either situation case, each conformation for unit *i* is annotated by an intrinsic energy value that is calculated using an empirical potential function of one's choosing. The conformations are binned into rotational isomeric states based on the similarities of the backbone and sidechain dihedral

angles (see Figure 1.1). Residue / unit *i*, might have *m* rotational isomers whereas residue *j* might have *n* rotational isomers. For a given residue, each rotational isomer is assigned a weight that is calculated using the Boltzmann weights of energies of individual conformations that make up the rotational isomer.

Given an amino acid sequence of *N* residues, one can calculate, *a priori*, the probabilities associated with all combinations rotational isomers. For the sequence of interest the number of rotational isomers per residue, their statistical weights, and the sequence composition dictates the total number of conformational possibilities and the likelihoods associated with each conformation. These likelihoods make up the predicted conformational distribution function and can be used to calculate a variety of conformational properties including the average end-to-end distance, the average radius of gyration, the average hydrodynamic size, the average distance between residues *i* and *j*, and any observable that can be cast as a function of a moment of the conformational distribution function.

An alternative approach is to analyze experimental data, specifically data from fluorescence [(85-87)] or force spectroscopy [(88-90)] that are functions of end-to-end distances using variants of the worm like chain model. The persistence length l_p is the length scale over which the chain behaves like a continuously deformable entity. For a rod-like chain l_p equals the contour length and for a freely-joined chain, l_p equals the bond length, so this model ostensibly allows interpolation between two extremes. Fluctuations are highly correlated for spatial separations that are smaller than l_p. For spatial separations longer than l_p, the worm-like segments become uncorrelated and the model reverts to the Flory random coil limit. Therefore, if l_p is found to be small, the worm like chain model does not yield any insights that go beyond the Flory random coil.

Estimates of l_p values for different sequences, studied under similar solution conditions yield comparative assessments of sequence-ensemble relationships through comparative measures of "chain stiffness" although Yamakawa [91] has highlighted the limitations of l_p as a measure of stiffness. The assumption of continuous deformation for spatial separations less than l_P is questionable because this assumption breaks down if the chain can form heterogeneous ensembles of compact globules. Despite its inherent weaknesses, the worm like chain model retains appeal for its ease of use in interpreting experimental data for IDPs and denatured proteins [85-87].

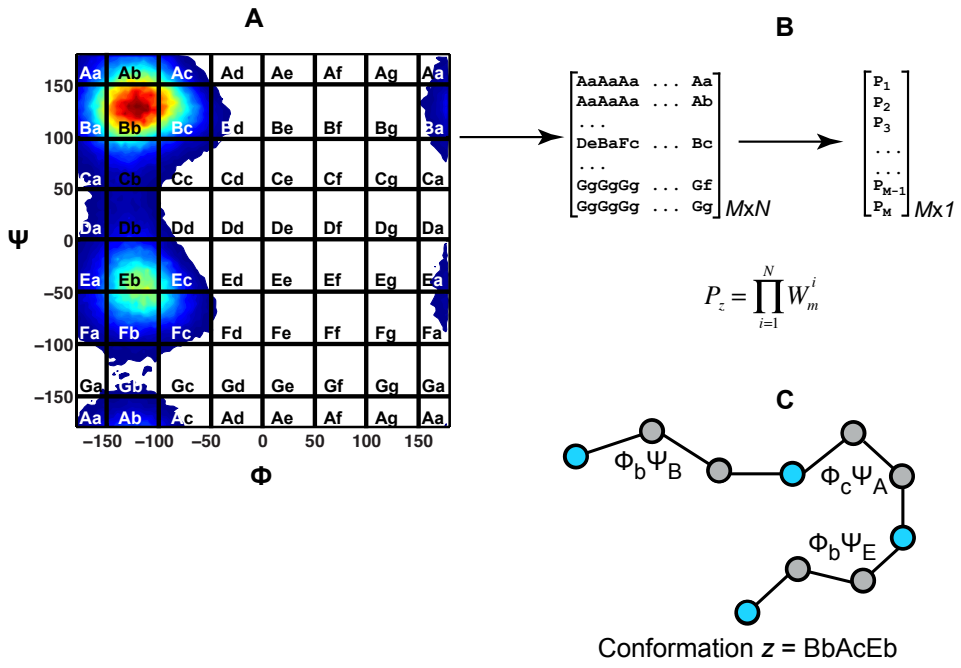


Figure 1.1: Illustration of how the rotational isomeric approximation of the Flory random coil model is constructed. (A) This process begins with a detailed calculation of the free energy landscape (with free energies increasing from red to blue) for an individual amino acid (or Kuhn segment), shown here for alanine. The tiles represent a coarse graining of conformational space into discrete rotational isomers, and each isomer has a label and a statistical weight that is calculated using the energies associated with conformations that make up a rotational isomer.

The assumption of independence / additivity allows the statistical weights for each combination of rotational isomers to be written as a product of individual weights. Panel (B) shows this procedure, whereby there are *M* conformations for a polypeptide of *N* residues and the statistical weight for each conformation *z* is a product of the weights for individual residues. The result is a weighted ensemble of all conformational possibilities where each "conformation" is denoted using a combination of the coarse grain rotational isomers. Panel (C) shows a schematic conformation for one of the conformations *z*.

1.7 The value of limiting models

The preceding discussion focuses on limiting models, which are analogous to limiting models / laws in other branches of physics that include the Debye-Hückel equation for calculating activity coefficients of electrolytes [(92)], and the Hildebrand [(93)] / Flory-Huggins expressions for the free energies of ideal mixtures [(94, 95)]. Limiting laws or models provide a route for interpreting experimental data as deviations from ideal behavior. As a limiting model, the Flory random coil model is often used to calibrate measured observables such as NMR chemical shifts [(96, 97)] and NMR paramagnetic relaxation enhancement effects [(45)] i.e., observables can be calibrated as deviations from the Flory random coil. This helps assess the contributions of spatial interactions between residues that are distal in the linear sequence. Such approaches are decidedly one-sided because deviations from a limiting model tell us what an ensemble is not and this is inadequate for developing a complete understanding of sequence-ensemble relationships.

1.8 Toward more realistic polymer models for IDPs

Conformational statistics are dictated by the interplay between chain-solvent and intrachain (intra-backbone, backbone-sidechain, and sidechain-sidechain) interactions [(82)]. As a

result, polymers undergo continuous transitions between distinct conformational classes. These transitions are modulated by changes to solvent-mediated interactions either through the addition of ternary components or by changing the temperature and / or pressure. Importantly, the conformational classes are akin to distinct phases because they have distinct density profiles and the variation of spatial separation as a function of sequence separation follows distinct patterns. Transitions between conformational classes are hence akin to phase transitions [(98)].

1.9 Measures for assessing the phase behavior of polymers

Quantities such as the average radius of gyration ($\langle R_g \rangle$), the average hydrodynamic radius ($\langle R_h \rangle$), and the average end-to-end distance ($\langle R_{ee} \rangle$), respectively are different measures of chain size that can be used to quantify the average density, intrinsic viscosity, and concentration of one end of the chain around the other. In addition to measures of chain size, one can also calculate the average shapes of polymers. The average asphericity (δ^*) quantifies the extent of deviation from a perfect sphere ($\delta^* = 0$). For ellipsoids, $\delta^* \approx 0.4$, and this quantity attains its maximum value of 1 for a perfect rod [(99)]. The average asphericity is calculated from the ensemble-averaged eigenvalues of the gyration tensor.

One can also calculate the average distances between residues i and j. The quantity $\langle R_{ij} \rangle$ represents the ensemble-average of spatial separations calculated as averages over all pairs i and j that yield a sequence separation |j–i| [(100)]. Multiple pairs of residues i and j will have similar sequence separations |j–i|. The profile of $\langle R_{ij} \rangle$ plotted against sequence separation |j–i| quantifies the local concentration of chain segments around each other and provides the most detailed information regarding the so-called link density [(101)], which is a formal order parameter in formalisms of polymer theories such as the Lifshitz approach [(100, 102-105)]. In addition to ensemble averages, one can also calculate the one- and two-parameter distribution functions such as $P(R_g)$,

$P(R_{ee})$, $P(R_h)$, $P(\delta)$, $P(R_{ij}|\,|j-i|)$, and $P(R_g, \delta)$. The latter quantifies the joint distribution of sizes and shapes.

Importantly, all of the quantities listed above are accessible to the appropriate combination of experiments and can be calculated using coordinates for simulated ensembles. This enables quantitative comparisons between simulation results and experiments thus facilitating direct approaches to either test predictions from simulations or routes to incorporate experimental data as restraints in simulations for generating ensembles that best describe experimental data. Both approaches are equally important and have enabled the development of quantitative sequence-ensemble relationships for IDPs.

1.10 Assessing solvent quality

The balance between chain-solvent interactions and intra-chain interactions is quantified using a parameter v_{ex}. This quantity has units of volume and is proportional to the integral of the Mayer *f*-function, *i.e.*, $v_{ex} = -\int f(r) d^3r$ where $f(r) = \exp\left[-\beta W(r)\right] - 1$; $W(r)$, is the potential of mean force for the thermally averaged inter-residue interaction and $\beta = 1/RT$ where R is Boltzmann's constant and T is temperature. If the effective inter-residue interactions are repulsive, then the Mayer-*f* function is negative, which leads to positive values for v_{ex} and the converse is true for inter-residue interactions that are attractive on average. The parameter v_{ex} is hence a measure of the volume excluded, per residue, for favorable interactions with the surrounding solvent that results from the competition between chain-chain and chain-solvent interactions. It provides a measure of the strengths of pairwise inter-residue interactions, on average, and can be related to the second virial coefficient that is accessible using light scattering measurements [(82)].

1.11 Classifying ensemble types based on v_{ex}

In a good solvent, $v_{ex} > 0$, and the chain expands to maximize the polymer-solvent interface. Expanded unfolded states are sampled *in vitro* in high concentrations of chemical denaturants such as urea and guanidinium chloride. Aqueous solutions with high concentrations (8 M) urea are presumed to be reasonable mimics of good solvents for generic polypeptides because urea, a carbonyl diamide, is chemically equivalent to polypeptide backbone amides. As a result, quantities such as $\langle R_g \rangle$ [106], $\langle R_h \rangle$ [107], and the average end-to-end distance $\langle R_{ee} \rangle$ scale as $N^{0.59}$ with chain length, N. In a good solvent the distances $\langle R_{ij} \rangle$ scale as $|j-i|^{0.59}$ as a function of sequence separation $|j-i|$.

Since the inter-residue interactions are repulsive on average, $v_{ex} > 0$ in good solvents. Indeed, the sizes of self-avoiding random walks also scale as $N^{0.59}$ and conformational ensembles for polymers in good solvents and self-avoiding random walks are said to belong to the same "universality class" [108]. Accordingly, ensembles generated in atomistic detail for proteins in the excluded volume (EV) limit are useful reference states for expanded unfolded states [109-115]. In the EV limit, ensembles are generated using atomistic descriptions of proteins and all non-bonded interactions excepting steric repulsions are ignored.

The low overall hydrophobicity of IDP sequences implicitly suggests that these systems come under the same rubric as chemically denatured proteins. Hence, a popular approach is to generate EV limit ensembles and filter out those conformations that cause deviations from observables that are measured experimentally [87, 116-122]. Although this seems like a reasonable approach, it imposes the fiat that aqueous solutions are mimic of good solvents for IDP sequences and ignores the possibility that these sequences can sample compact phases.

1.12 Implications of decreasing values of v_{ex}

The parameter v_{ex} can change continuously going from positive values in a good solvent, through zero in a theta solvent, to negative values in a poor solvent [82]. If the effects of chain-solvent and intra-chain interactions exactly counterbalance, then $v_{ex}=0$, and the chain is said to be in a theta solvent. Under such conditions, the chain statistics are consistent with those of a Flory random coil model. It is important to note that this behavior comes about due to counterbalancing of interactions rather than explicitly ignoring non-local interactions. In a theta solvent, $\langle R_g \rangle$, $\langle R_h \rangle$, and $\langle R_{ee} \rangle$ scale as $N^{0.5}$ and $\langle R_{ij} \rangle$ scales as $|j-i|^{0.5}$.

In a poor solvent, $v_{ex} < 0$ and the chain prefers compact, globular conformations that minimize the polymer-solvent interface and $\langle R_g \rangle$ and $\langle R_h \rangle$ scale as $N^{1/3}$ with chain length, N. The sizes of folded proteins follow $N^{1/3}$ scaling [123]. The poorer the solvent, the more negative the value of v_{ex}. Statistics of inter-residue distances change fundamentally in a poor solvent. The distances $\langle R_{ij} \rangle$ do not increase with sequence separation $|j-i|$ according to a power law. Instead, for all values of $|j-i|$ greater than larger than a so-called blob length (ca. 5-7 residues), the value of $\langle R_{ij} \rangle$ is fixed by the average density of the globule.

1.13 Layout of the thesis

In the chapters that follow the focus is on characterizing the degree and nature of conformational heterogeneity in a series of systems that are directly relevant to the problems of protein folding, protein stability, requirements for structure in seeding incipient associations between physiologically relevant IDPs, and the modulation of disorder and associativity of IDPs by sequence composition and sequence context. Instead of taking a generic, one approach fits all systems route we use experimental observations that raise questions about archetypes of the two categories of IDPs. We use molecular simulations combined with analyses that are motivated by polymer physics theories and physical theories for self-assembling systems to provide answers to

questions / puzzles that are raised by experimental observations. This approach has the desired effect of directly addressing biologically relevant issues while providing quantitative insights regarding the options afforded by the multitude of conformational options available to both categories of IDPs. Having presented salient results from analyses for experimentally motivated systems, we end with a description of recent work that introduces a new parameter that is designed to quantify the degree of conformational heterogeneity within an ensemble. Its importance is highlighted by demonstrating the insights obtained for different systems, especially when used jointly with descriptors of heterogeneity that are derived from polymer physics theories.

1.14 References

1. Fischer, E. (1894) Einfuss der conguration auf die wirkung der enzyme, *Berichte der deutschen chemischen Gesellschaft 27*, 2985-2993.
2. Fischer, E. (1894) Einfluss der conguration auf die wirkung der enzyme. II, *Berichte der deutschen chemischen Gesellschaft 27*, 3479-3483.
3. Fischer, E. (1898) Bedeutung der Stereochemie fur die Physiologie, *Zeitschrift fur Physiologische Chemie 26*, 60-87.
4. Koshland, D. E. (1995) The key-lock theory and the induced fit theory, *Angewandte Chemie-International Edition in English 33*, 2375-2378.
5. Monod, J., Wyman, J., and Changeux, J.-P. (1964) On the nature of allosteric transitions: a plausible model, *Journal of Molecular Biology 12*, 88-118.
6. Yu, E. W., and Koshland, D. E. (2001) Propagating conformational changes over long (and short) distances in proteins, *Proceedings of the National Academy of Sciences of the United States of America 98*, 9517-9520.

7. Uversky, V. N. (2002) Natively unfolded proteins: A point where biology waits for physics, *Protein Science 11*, 739-756.

8. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., and Obradovic, Z. (2001) Intrinsically disordered protein, *J Mol Graph Model 19*, 26-59.

9. Wright, P. E., and Dyson, H. J. (1999) Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm, *Journal of Molecular Biology 293*, 321-331.

10. Kriwacki, R. W., Hengst, L., Tennant, L., Reed, S. I., and Wright, P. E. (1996) Structural studies of p21(Waf1/Cip1/Sdi1) in the free and Cdk2-bound state: Conformational disorder mediates binding diversity, *Proceedings of the National Academy of Sciences of the United States of America 93*, 11504-11509.

11. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., and Obradovic, Z. (2002) Intrinsic disorder and protein function, *Biochemistry 41*, 6573-6582.

12. Dunker, A. K., Brown, C. J., and Obradovic, Z. (2002) Identification and functions of usefully disordered proteins. , *Advances in Protein Chemistry 62*, 25-49.

13. Dyson, H. J., and Wright, P. E. (2005) Intrinsically unstructured proteins and their functions, *Nature Reviews in Molecular Cell Biology 6*, 197-208.

14. Tompa, P. (2003) The functional benefits of protein disorder, *Journal Of Molecular Structure-Theochem 666*, 361-371.

15. Uversky, V. N., Gillespie, J. R., and Fink, A. L. (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions?, *Proteins-Structure Function And Genetics 41*, 415-427.

16. Dyson, H. J., and Wright, P. E. (2002) Coupling of folding and binding for unstructured proteins, *Current Opinion in Structural Biology 12*, 54-60.

17. Frankel, A. D., and Smith, C. A. (1998) Induced folding in RNA-protein recognition: More than a simple molecular handshake, *Cell 92*, 149-151.

18. Mucsi, Z., Hudecz, F., Hollosi, M., Tompa, P., and Friedrich, P. (2003) Binding-induced folding transitions in calpastatin subdomains A and C, *Protein Science 12*, 2327-2336.

19. Lacy, E. R., Filippov, I., Lewis, W. S., Otieno, S., Xiao, L. M., Weiss, S., Hengst, L., and Kriwacki, R. W. (2004) p27 binds cyclin-CDK complexes through a sequential mechanism involving binding-induced protein folding, *Nature Structural & Molecular Biology 11*, 358-364.

20. Receveur-Brechot, V., Bourhis, J. M., Uversky, V. N., Canard, B., and Longhi, S. (2006) Assessing protein disorder and induced folding, *Proteins-Structure Function And Bioinformatics 62*, 24-45.

21. Kohler, J. J., Metallo, S. J., Schneider, T. L., and Schepartz, A. (1999) DNA specificity enhanced by sequential binding of protein monomers, *Proceedings of the National Academy of Sciences, USA 96*, 11735-11739.

22. Liu, J. G., Perumal, N. B., Oldfield, C. J., Su, E. W., Uversky, V. N., and Dunker, A. K. (2006) Intrinsic disorder in transcription factors, *Biochemistry 45*, 6873-6888.

23. Fuxreiter, M., Tompa, P., Simon, I., Uversky, V. N., Hansen, J. C., and Asturias, F. J. (2008) Malleable machines take shape in eukaryotic transcriptional regulation, *Nature Chemical Biology 4*, 728-737.

24. Spolar, R. S., and Record, M. T. (1994) Coupling of local folding to site-specific binding of proteins to DNA, *Science 263*, 777-784.

25. von Hippel, P. H. (2007) From "simple" DNA-protein interactions to the macromolecular machines of gene expression, In *Annual Review of Biophysics and Biomolecular Structure*, pp 79-105.

26. Tompa, P., and Fuxreiter, M. (2008) Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions, *Trends In Biochemical Sciences 33*, 2-8.

27. Fuxreiter, M., and Tompa, P. (2009) Fuzzy interactome: the limitations of models in molecular biology, *Trends In Biochemical Sciences 34*, 3-3.

28. Mittag, T., Marsh, J., Orlicky, S., Borg, M., Tang, X., Sicheri, F., Chan, H. S., Kay, L. E., Tyers, M., and Forman-Kay, J. D. (2010) "Fuzzy" complexes: How much disorder can a biologically relevant complex tolerate, and can it even be beneficial?, *Biochemistry and Cell Biology-Biochimie et Biologie Cellulaire 88*, 403-403.

29. Padrick, S. B., and Miranker, A. D. (2001) Islet amyloid polypeptide: Identification of long-range contacts and local order on the fibrillogenesis pathway, *Journal of Molecular Biology 308*, 783-794.

30. Fandrich, M., and Dobson, C. M. (2002) The behaviour of polyamino acids reveals an inverse side chain effect in amyloid structure formation, *Embo J 21*, 5682-5690.

31. Bitan, G., Kirkitadze, M. D., Lomakin, A., Vollers, S. S., Benedek, G. B., and Teplow, D. B. (2003) Amyloid beta-protein (Abeta) assembly: Abeta40 and Abeta42 oligomerize through distinct pathways, *Proc. Natl. Acad. Sci. U. S. A. 100*, 330-335.

32. Scheibel, T., Bloom, J., and Lindquist, S. L. (2004) The elongation of yeast prion fibers involves separable steps of association and conversion, *Proceedings of the National Academy of Sciences of the United States of America 101*, 2287-2292.

33. Uversky, V. N., and Fink, A. L. (2004) Conformational constraints for amyloid fibrillation: The importance of being unfolded, *Biochimica et Biophysica Acta - Proteins and Proteomics 1698*, 131.

34. Calamai, M., Chiti, F., and Dobson, C. M. (2005) Amyloid fibril formation can proceed from different conformations of a partially unfolded protein, *Biophysical Journal 89*, 4201.

35. Krishnan, R., and Lindquist, S. L. (2005) Structural insights into a yeast prion illuminate nucleation and strain diversity, *Nature 435*, 765-772.

36. Halfmann, R., Alberti, S., Krishnan, R., Lyle, N., O'Donnell, C. W., King, O. D., Berger, B., Pappu, R. V., and Lindquist, S. (2011) Opposing Effects of Glutamine and Asparagine Govern Prion Formation by Intrinsically Disordered Proteins, *Molecular Cell 43*, 72-84.

37. Heim, M., Romer, L., and Scheibel, T. (2010) Hierarchical structures made of proteins. The complex architecture of spider webs and their constituent silk proteins, *Chemical Society Reviews 39*, 156-164.

38. Vitalis, A., and Pappu, R. V. (2011) Assessing the contribution of heterogeneous distributions of oligomers to aggregation mechanisms of polyglutamine peptides, *Biophysical Chemistry 159*, 14-23.

39. Weathers, E. A., Paulaitis, M. E., Woolf, T. B., and Hoh, J. H. (2004) Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein, *FEBS Letters 576*, 348-352.

40. Gsponer, J., Futschik, M. E., Teichmann, S. A., and Babu, M. M. (2008) Tight Regulation of Unstructured Proteins: From Transcript Synthesis to Protein Degradation, *Science 322*, 1365-1368.

41. Tsvetkov, P., Reuven, N., and Shaul, Y. (2009) The nanny model for IDPs, *Nature Chemical Biology 5*, 778-781.

42. Tsvetkov, P., Asher, G., Paz, A., Reuven, N., Sussman, J. L., Silman, I., and Shaul, Y. (2008) Operational definition of intrinsically unstructured protein sequences based on susceptibility to the 20S proteasome, *Proteins-Structure Function And Bioinformatics 70*, 1357-1366.

43. Eliezer, D., Kutluay, E., Bussell, R., and Browne, G. (2001) Conformational properties of alpha-synuclein in its free and lipid-associated states, *Journal of Molecular Biology 307*, 1061-1073.

44. Dyson, H. J., and Wright, P. E. (2002) Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance, *Adv Protein Chem 62*, 311-340.

45. Dyson, H. J., and Wright, P. E. (2004) Unfolded proteins and protein folding studied by NMR, *Chemical Reviews 104*, 3607-3622.

46. Barre, P., and Eliezer, D. (2006) Folding of the repeat domain of tau upon binding to lipid surfaces, *Journal of Molecular Biology 362*, 312-326.

47. Mittag, T., and Forman-Kay, J. D. (2007) Atomic-level characterization of disordered protein ensembles, *Current Opinion In Structural Biology 17*, 3-14.

48. Bezsonova, I., Forman-Kay, J., and Prosser, R. S. (2008) Molecular oxygen as a paramagnetic NMR probe of protein solvent exposure and topology, *Concepts in Magnetic Resonance Part A 32A*, 239-253.

49. Jensen, M. R., Markwick, P. R. L., Meier, S., Griesinger, C., Zweckstetter, M., Grzesiek, S., Bernado, P., and Blackledge, M. (2009) Quantitative Determination of the Conformational Properties of Partially Folded and Intrinsically Disordered Proteins Using NMR Dipolar Couplings, *Structure 17*, 1169-1185.

50. Salmon, L., Nodet, G., Ozenne, V., Yin, G. W., Jensen, M. R., Zweckstetter, M., and Blackledge, M. (2010) NMR Characterization of Long-Range Order in Intrinsically Disordered Proteins, *Journal of the American Chemical Society 132*, 8407-8418.

51. Mukhopadhyay, S., Krishnan, R., Lemke, E. A., Lindquist, S., and Deniz, A. A. (2007) A natively unfolded yeast prion monomer adopts an ensemble of collapsed and rapidly fluctuating structures, *Proceedings of the National Academy of Sciences of the United States of America 104*, 2649-2654.

52. Ferreon, A. C. M., Gambin, Y., Lemke, E. A., and Deniz, A. A. (2009) Interplay of alpha-synuclein binding and conformational switching probed by single-molecule fluorescence, *Proceedings of the National Academy of Sciences of the United States of America 106*, 5645-5650.

53. Nettels, D., Muller-Spath, S., Kuster, F., Hofmann, H., Haenni, D., Ruegger, S., Reymond, L., Hoffmann, A., Kubelka, J., Heinz, B., Gast, K., Best, R. B., and Schuler, B. (2009) Single-molecule spectroscopy of the temperature-induced collapse of unfolded proteins, *Proceedings of the National Academy of Sciences of the United States of America 106*, 20740-20745.

54. Muller-Spath, S., Soranno, A., Hirschfeld, V., Hofmann, H., Ruegger, S., Reymond, L., Nettels, D., and Schuler, B. (2010) Charge interactions can dominate the dimensions of intrinsically disordered proteins, *Proceedings of the National Academy of Sciences of the United States of America 107*, 14609-14614.

55. Rao, J. N., Jao, C. C., Hegde, B. G., Langen, R., and Ulmer, T. S. (2010) A Combinatorial NMR and EPR Approach for Evaluating the Structural Ensemble of Partially Folded Proteins, *Journal of the American Chemical Society 132*, 8657-8668.

56. Paz, A., Zeev-Ben-Mordehai, T., Lundqvist, M., Sherman, E., Mylonas, E., Weiner, L., Haran, G., Svergun, D. I., Mulder, F. A. A., Sussman, J. L., and Silman, I. (2008) Biophysical characterization of the unstructured cytoplasmic domain of the human neuronal adhesion protein neuroligin 3, *Biophysical Journal 95*, 1928-1944.

57. Wells, M., Tidow, H., Rutherford, T. J., Markwick, P., Jensen, M. R., Mylonas, E., Svergun, D. I., Blackledge, M., and Fersht, A. R. (2008) Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain, *Proceedings of the National Academy of Sciences of the United States of America 105*, 5762-5767.

58. Jensen, M. R., Salmon, L., Nodet, G., and Blackledge, M. (2010) Defining Conformational Ensembles of Intrinsically Disordered and Partially Folded Proteins Directly from Chemical Shifts, *Journal of the American Chemical Society 132*, 1270-+.

59. Zhang, W. H., Ganguly, D., and Chen, J. H. (2012) Residual Structures, Conformational Fluctuations, and Electrostatic Interactions in the Synergistic Folding of Two Intrinsically Disordered Proteins, *Plos Computational Biology 8*.

60. Ganguly, D., Zhang, W. H., and Chen, J. H. (2012) Synergistic folding of two intrinsically disordered proteins: searching for conformational selection, *Molecular BioSystems 8*, 198-209.

61. Ganguly, D., and Chen, J. H. (2009) Atomistic Details of the Disordered States of KID and pKID. Implications in Coupled Binding and Folding, *Journal of the American Chemical Society 131*, 5214-5223.

62. De Sancho, D., and Best, R. B. (2012) Modulation of an IDP binding mechanism and rates by helix propensity and non-native interactions: association of HIF1 alpha with CBP, *Molecular BioSystems 8*, 256-267.

63. Espinoza-Fonseca, L. M., Ilizaliturri-Flores, I., and Correa-Basurto, J. (2012) Backbone conformational preferences of an intrinsically disordered protein in solution, *Molecular BioSystems 8*, 1798-1805.

64. Moritsugu, K., Terada, T., and Kidera, A. (2012) Disorder-to-Order Transition of an Intrinsically Disordered Region of Sortase Revealed by Multiscale Enhanced Sampling, *Journal of the American Chemical Society 134*, 7094-7101.

65. Vitalis, A., Wang, X., and Pappu, R. V. (2007) Quantitative Characterization of Intrinsic Disorder in Polyglutamine: Insights from Analysis Based on Polymer Theories, *Biophys. J. 93*, 1923-1937.

66. Vitalis, A., and Caflisch, A. (2010) Micelle-Like Architecture of the Monomer Ensemble of Alzheimer's Amyloid-beta Peptide in Aqueous Solution and Its Implications for A beta Aggregation, *Journal of Molecular Biology 403*, 148-165.

67. Wostenberg, C., Kumar, S., Noid, W. G., and Showalter, S. A. (2011) Atomistic Simulations Reveal Structural Disorder in the RAP74-FCP1 Complex, *Journal of Physical Chemistry B 115*, 13731-13739.

68. Lindorff-Larsen, K., Best, R. B., DePristo, M. A., Dobson, C. M., and Vendruscolo, M. (2005) Simultaneous determination of protein structure and dynamics, *Nature 433*, 128-132.

69. Vendruscolo, M. (2007) Determination of conformationally heterogeneous states of proteins, *Current Opinion In Structural Biology 17*, 15-20.

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

71. Robustelli, P., Kohlhoff, K., Cavalli, A., and Vendruscolo, M. (2010) Using NMR Chemical Shifts as Structural Restraints in Molecular Dynamics Simulations of Proteins, *Structure 18*, 923-933.

72. De Simone, A., Montalvao, R. W., and Vendruscolo, M. (2011) Determination of Conformational Equilibria in Proteins Using Residual Dipolar Couplings, *Journal Of Chemical Theory And Computation 7*, 4189-4195.

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

74. Markwick, P. R. L., Cervantes, C. F., Abel, B. L., Komives, E. A., Blackledge, M., and McCammon, J. A. (2010) Enhanced Conformational Space Sampling Improves the Prediction of Chemical Shifts in Proteins, *Journal of the American Chemical Society 132*, 1220-+.

75. Choy, W. Y., and Forman-Kay, J. D. (2001) Calculation of ensembles of structures representing the unfolded state of an SH3 domain, *J Mol Biol 308*, 1011-1032.

76. Marsh, J. A., and Forman-Kay, J. D. (2012) Ensemble modeling of protein disordered states: Experimental restraint contributions and validation, *Proteins-Structure Function And Bioinformatics 80*, 556-572.

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

78. Lo Conte, L., Ailey, B., Hubbard, T. J. P., Brenner, S. E., Murzin, A. G., and Chothia, C. (2000) SCOP: a Structural Classification of Proteins database, *Nucleic Acids Research 28*, 257-259.

79. Nagano, N., Orengo, C. A., and Thornton, J. M. (2002) One fold with many functions: The evolutionary relationships between TIM barrel families based on their sequences, structures and functions, *Journal of Molecular Biology 321*, 741-765.

80. Orengo, C. A., Bray, J. E., Buchan, D. W. A., Harrison, A., Lee, D., Pearl, F. M. G., Sillitoe, I., Todd, A. E., and Thornton, J. M. (2002) The CATH protein family database: A resource for structural and functional annotation of genomes, *Proteomics 2*, 11-21.

81. Rison, S. C. G., Teichmann, S. A., and Thornton, J. M. (2002) Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in Escherichia coli, *Journal of Molecular Biology 318*, 911-932.

82. Rubinstein, M., and Colby, R. H. (2003) *Polymer Physics*, Oxford University Press, Oxford and New York.

83. Flory, P. J. (1969) *Statistical Mechanics of Chain Molecules*, Oxford University Press, New York.

84. Yamakawa, H. (1976) Statisical mechanics of wormlike chains, *Pure and Applied Chemistry 46*, 135-141.

85. Lapidus, L. J., Steinbach, P. J., Eaton, W. A., Szabo, A., and Hofrichter, J. (2002) Effects of chain stiffness on the dynamics of loop formation in polypeptides. Appendix: Testing a 1-dimensional diffusion model for peptide dynamics, *Journal of Physical Chemistry B 106*, 11628-11640.

86. Buscaglia, M., Lapidus, L. J., Eaton, W. A., and Hofrichter, J. (2006) Effects of denaturants on the dynamics of loop formation in polypeptides, *Biophysical Journal 91*, 276-288.

87. Singh, V. R., and Lapidus, L. J. (2008) The intrinsic stiffness of polyglutamine peptides, *Journal of Physical Chemistry B 112*, 13172.

88. Kellermayer, M. S. Z., Smith, S. B., Granzier, H. L., and Bustamante, C. (1997) Folding-unfolding transitions in single titin molecules characterized with laser tweezers, *Science 276*, 1112-1116.

89. Bemis, J. E., Akhremitchev, B. B., and Walker, G. C. (1999) Single polymer chain elongation by atomic force microscopy, *Langmuir 15*, 2799-2805.

90. Bright, J. N., Woolf, T. B., and Hoh, J. H. (2001) Predicting properties of intrinsically unstructured proteins, *Prog. Biophys. Mol. Biol. 76*, 131-173.

91. Yamakawa, H. (1997) *Helical wormlike chains in polymer solutions*, Springer, New York.

92. Debye, P., and Huckel, E. (1923) Zur Theorie der Elektrolyte, *Physikalische Zeitschrift 24*, 185-206.

93. Hildebrand, J. H. (1923) Theory of solubility, *Physical Review 21*, 46-52.

94. Huggins, M., L. (1941) Solutions of long chain compounds, *Journal of Chemical Physics 9*, 440-440.

95. Flory, P. J. (1942) Thermodynamics of high polymer solutions, *Journal of Chemical Physics 10*, 51-61.

96. Schwarzinger, S., Kroon, G. J. A., Foss, T. R., Wright, P. E., and Dyson, H. J. (2000) Random coil chemical shifts in acidic 8 M urea: Implementation of random coil shift data in NMRView, *Journal Of Biomolecular Nmr 18*, 43-48.

97. Schwarzinger, S., Kroon, G. J. A., Foss, T. R., Chung, J., Wright, P. E., and Dyson, H. J. (2001) Sequence-dependent correction of random coil NMR chemical shifts, *Journal Of The American Chemical Society 123*, 2970-2978.

98. de Gennes, P.-G. (1979) *Scaling Concepts in Polymer Physics*, Cornell University Press, Ithaca and London.

99. Steinhauser, M. O. (2005) A molecular dynamics study on universal properties of polymer chains in different solvent qualities. Part I. A review of linear chain properties, *Journal of Chemical Physics 122*.

100. Imbert, J. B., Lesne, A., and Victor, J. M. (1997) Distribution of the order parameter of the coil-globule transition, *Physical Review E 56*, 5630-5647.

101. Lifshitz, I. M., Grosberg, A.Y., Khokhlov, A.R. . (1978) Some problems of the statistical physics of polymer chains with volume interaction, *Reviews of Modern Physics 50*, 683-713.

102. Grosberg, A. Y., and Kuznetsov, D. V. (1992) Quantitative Theory Of The Globule-To-Coil Transition .1. Link Density Distribution In A Globule And Its Radius Of Gyration, *Macromolecules 25*, 1970-1979.

103. Grosberg, A. Y., and Kuznetsov, D. V. (1992) Quantitative Theory Of The Globule-To-Coil Transition .2. Density Density Correlation In A Globule And The Hydrodynamic Radius Of A Macromolecule, *Macromolecules 25*, 1980-1990.

104. Grosberg, A. Y., and Kuznetsov, D. V. (1992) Quantitative Theory Of The Globule-To-Coil Transition .3. Globule Globule Interaction And Polymer-Solution Binodal And Spinodal Curves In The Globular Range, *Macromolecules 25*, 1991-1995.

105. Grosberg, A. Y., and Kuznetsov, D. V. (1992) Quantitative Theory Of The Globule-To-Coil Transition .4. Comparison Of Theoretical Results With Experimental-Data, *Macromolecules 25*, 1996-2003.

106. Kohn, J. E., I.S. Millett, J. Jacob, B. Zagrovic, T.M. Dillon, N. Cingel, R.S. Dothager, S. Seifert, P. Thiyagarajan, T.R. Sosnick, M.Z. Hasan, V.S. Pande, I. Ruzcinski, S. Doniach, and Plaxco, K. W. (2004) Random-coil behavior and the dimensions of chemically unfolded proteins, *Proceedings of the National Academy of Sciences of the United States of America 101*, 12491-12496.

107. Penkett, C. J., Redfield, C., Dodd, I., Hubbard, J., McBay, D. L., Mossakowska, D. E., Smith, R. A., Dobson, C. M., and Smith, L. J. (1997) NMR analysis of main-chain conformational preferences in an unfolded fibronectin-binding protein, *J Mol Biol 274*, 152-159.

108. Schäfer, L. (1999) *Excluded Volume Effects in Polymer Solutions as Explained by the Renormalization Group*, Springer, Berlin.

109. Zhou, H. X. (2004) Polymer models of protein stability, folding, and interactions, *Biochemistry 43*, 2141-2154.

110. Tran, H. T., Wang, X., and Pappu, R. V. (2005) Reconciling observations of sequence-specific conformational propensities with the generic polymeric behavior of denatured proteins, *Biochemistry 44*, 11369-11380.

111. Tran, H. T., and Pappu, R. V. (2006) Toward an accurate theoretical framework for describing ensembles for proteins under strongly denaturing conditions, *Biophysical Journal 91*, 1868-1886.

112. Jha, A. K., Colubri, A., Freed, K. F., and Sosnick, T. R. (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*, 13099-13104.

113. Ding, F., Jha, R. K., and Dokholyan, N. V. (2005) Scaling behavior and structure of denatured proteins, *Structure 13*, 1047-1054.

114. Fitzkee, N. C., and Rose, G. D. (2004) Reassessing random-coil statistics in unfolded proteins, *Proceedings of the National Academy of Sciences of the United States of America 101*, 12497-12502.

115. Bernado, P., and Blackledge, M. (2009) A Self-Consistent Description of the Conformational Behavior of Chemically Denatured Proteins from NMR and Small Angle Scattering, *Biophysical Journal 97*, 2839-2845.

116. Moglich, A., Joder, K., and Kiefhaber, T. (2006) End-to-end distance distributions and intrachain diffusion constants in unfolded polypeptide chains indicate intramolecular hydrogen bond formation, *Proceedings Of The National Academy Of Sciences Of The United States Of America 103*, 12394-12399.

117. Goldenberg, D. P. (2003) Computational simulation of the statistical properties of unfolded proteins, *J Mol Biol 326*, 1615-1633.

118. Wang, Y., Trewhelia, J., and Goldenberg, D. P. (2008) Small-angle x-ray scattering of reduced ribonuclease A: Effects of solution conditions and comparisons with a computational model of unfolded proteins, *Journal of Molecular Biology 377*, 1576-1592.

119. Johansen, D., Jeffries, C. M. J., Hammouda, B., Trewhella, J., and Goldenberg, D. P. (2011) Effects of Macromolecular Crowding on an Intrinsically Disordered Protein Characterized by Small-Angle Neutron Scattering with Contrast Matching, *Biophysical Journal 100*, 1120-1128.

120. Johansen, D., Trewhella, J., and Goldenberg, D. P. (2011) Fractal dimension of an intrinsically disordered protein: Small-angle X-ray scattering and computational study of the bacteriophage lambda N protein, *Protein Science 20*, 1955-1970.

121. Sziegat, F., Silvers, R., Hahnke, M., Jensen, M. R., Blackledge, M., Wirmer-Bartoschek, J., and Schwalbe, H. (2012) Disentangling the Coil: Modulation of Conformational and

Dynamic Properties by Site-Directed Mutation in the Non-Native State of Hen Egg White Lysozyme, *Biochemistry 51*, 3361-3372.

122. Schneider, R., Huang, J. R., Yao, M. X., Communie, G., Ozenne, V., Mollica, L., Salmon, L., Jensen, M. R., and Blackledge, M. (2012) Towards a robust description of intrinsic protein disorder using nuclear magnetic resonance spectroscopy, *Molecular BioSystems 8*, 58-68.

123. Dima, R. I., and Thirumalai, D. (2004) Asymmetry in the shapes of folded and denatured states of proteins, *Journal of Physical Chemistry B 108*, 6564-6570.

Chapter 2

Experiments and simulations show how long-range contacts can form in expanded unfolded proteins with negligible secondary structure

This chapter is adapted from an article [1] published in the Proceedings of the National Academy of Sciences. Wenli Meng (WM) and Daniel P. Raleigh (DPR) designed the experiments and WM conducted the experiments. Bowu Luan (BL) contributed reagents for these experiments. Nicholas J. Lyle (NJL), the candidate, and Rohit V. Pappu (RVP) developed a quantitative framework to interpret the experimental results in a broader context. NL performed the simulations and analyzed the results. This work was supported by NSF grants MCB-0919860 to DPR and MCB-1121867 to RVP.

2.1 Introduction

Quantitative descriptions of unfolded proteins are important for understanding collapse transitions [2], protein folding mechanisms [3], misfolding, aggregation [4, 5], and the effects of macromolecular crowding on protein stability [6, 7]. Expanded unfolded states are sampled in high concentrations of chemical denaturants such as urea and guanidinium chloride. The sizes of these denatured proteins, quantified using hydrodynamic radii ($\langle R_h \rangle$) or radii of gyration ($\langle R_g \rangle$), scale as $N^{0.59}$ with chain length [8-10]. This $N^{0.59}$ scaling arises because denatured proteins expand to make favorable contacts with the surrounding solvent implying that high concentrations of denaturants are good solvents for generic proteins.

In good solvents, the ensemble-averaged inter-residue pair interaction coefficient is

positive suggesting that the preference for favorable chain-solvent interactions leads to intra-chain interactions being repulsive on average [11]. The validity of this inference for denatured proteins has been demonstrated recently using a combination of single molecule experiments and polymer theory [12]. Quantitative descriptions of chain statistics for polymers in good solvents rely on the so-called excluded volume (EV) limit as an important reference state and this is true for denatured state ensembles (DSEs) as well. EV limit ensembles are typically generated using atomistic descriptions of proteins and ignoring all non-bonded interactions excepting steric repulsions [13-16]. Hence, the ensemble-averaged inter-residue interaction coefficient (related to the second virial coefficient) is, by construction, positive in the EV limit thus affording the reproduction of $N^{0.59}$ scaling [11]. Descriptions of chain statistics based on EV limit ensembles are routinely used as reference states for intrinsically disordered proteins (IDPs) and denatured proteins and have proven useful in interpreting the results of nuclear magnetic resonance (NMR) and small angle X-ray scattering (SAXS) measurements for these systems [13-22].

If measured properties of DSEs can be adequately explained using EV limit ensembles, then the implication is as follows: In a typical Flory-like mean field description, the energy for each polymer configuration can be written as, $U = U_{\text{EV}} + U_{\text{non-EV-intrachain}} + U_{\text{chain-solvent}}$ [23]. The EV limit is reached if the non-EV-intrachain interactions are exactly counterbalanced by chain-solvent interactions. Deviations from the EV limit arise if there is imperfect compensation between non-EV-intrachain interactions and chain-solvent interactions. If imperfect compensation results from attractive non-EV-intrachain interactions, then deviations from the EV limit can lead to persistent local structure and global compaction with deviation from $N^{0.59}$ scaling as has been observed for proteins under mild denaturing conditions [24-34] and for proteins under folding (*i.e.*, poor solvent [35]) conditions [36, 37]. Imperfect compensation that results from

long-range intramolecular electrostatic repulsions and / or favorable chain-solvent interactions will also cause deviations from $N^{0.59}$ scaling. In these cases, the deviation causes increased chain expansion as has been observed recently for highly charged IDPs in aqueous solutions and in the presence of denaturant [19, 38].

Of interest is the possibility that imperfect compensation can be achieved between non-EV-intrachain interactions and chain-solvent interactions without causing deviations from $N^{0.59}$ scaling. Although this is the intuitive and canonical expectation given the documented denatured state effects on protein folding [7], no quantitative evidence has been offered in support of imperfect compensation being achieved while preserving $N^{0.59}$ scaling. Here, we report a new archetypal dataset for the DSE of the N-terminal domain of the L9 (NTL9) ribosomal protein in 8.3 M urea showing chain expansion consistent with $N^{0.59}$ scaling and the presence of detectable long-range contacts despite negligible secondary structure. Paramagnetic relaxation enhancements (PREs) provide support for the presence of transient long-range contacts in the DSE of NTL9. *The PRE data cannot be explained using EV limit ensembles and hence highlights the presence of imperfect compensation with $N^{0.59}$ scaling*. To understand the observations for the DSE of NTL9 we used atomistic thermal unfolding simulations of NTL9. In the simulations, the temperature acts as a proxy for the effect of denaturant (which changes solvent quality) by modulating the balance between the preference for compact conformations (seen at low temperatures) and expanded conformations (seen at high temperatures). We identified a set of high temperatures whose ensembles yield averages that are congruent with experimental data for the DSE of NTL9 in 8.3 M urea. Analysis of these ensembles provides a quantitative reasoning for the observed properties of NTL9's DSE in 8.3 M urea. Our simulation approach does not impose any *a priori* assumptions of equivalence between thermal and urea denaturation. Instead,

we use it to identify unfolded ensembles that yield properties consistent with the full panel of measurements, and thus seek a quantitative, albeit computationally tractable framework for interpreting the experimental observations.

2.2 NTL9 samples predominantly expanded conformations with negligible secondary structure in its DSE in 8.3 M urea

The values of R_h and R_g were measured using NMR pulsed field gradient diffusion and SAXS experiments, respectively. The measured R_h for NTL9 DSE in 8.3 M urea at 12 ˚C is 22.5 Å, and the measured R_g is 21.3 ± 1.5 Å at a protein concentration of 7.5 mg/mL as shown in Figure 2.1 and Figure 2.2. At a similar concentration, the measured R_g value for the native state is 12.2 Å as detailed in Figure 2.3. The measures R_h and R_g values are consistent with reported scaling laws for highly unfolded polypeptide chains [9, 10].

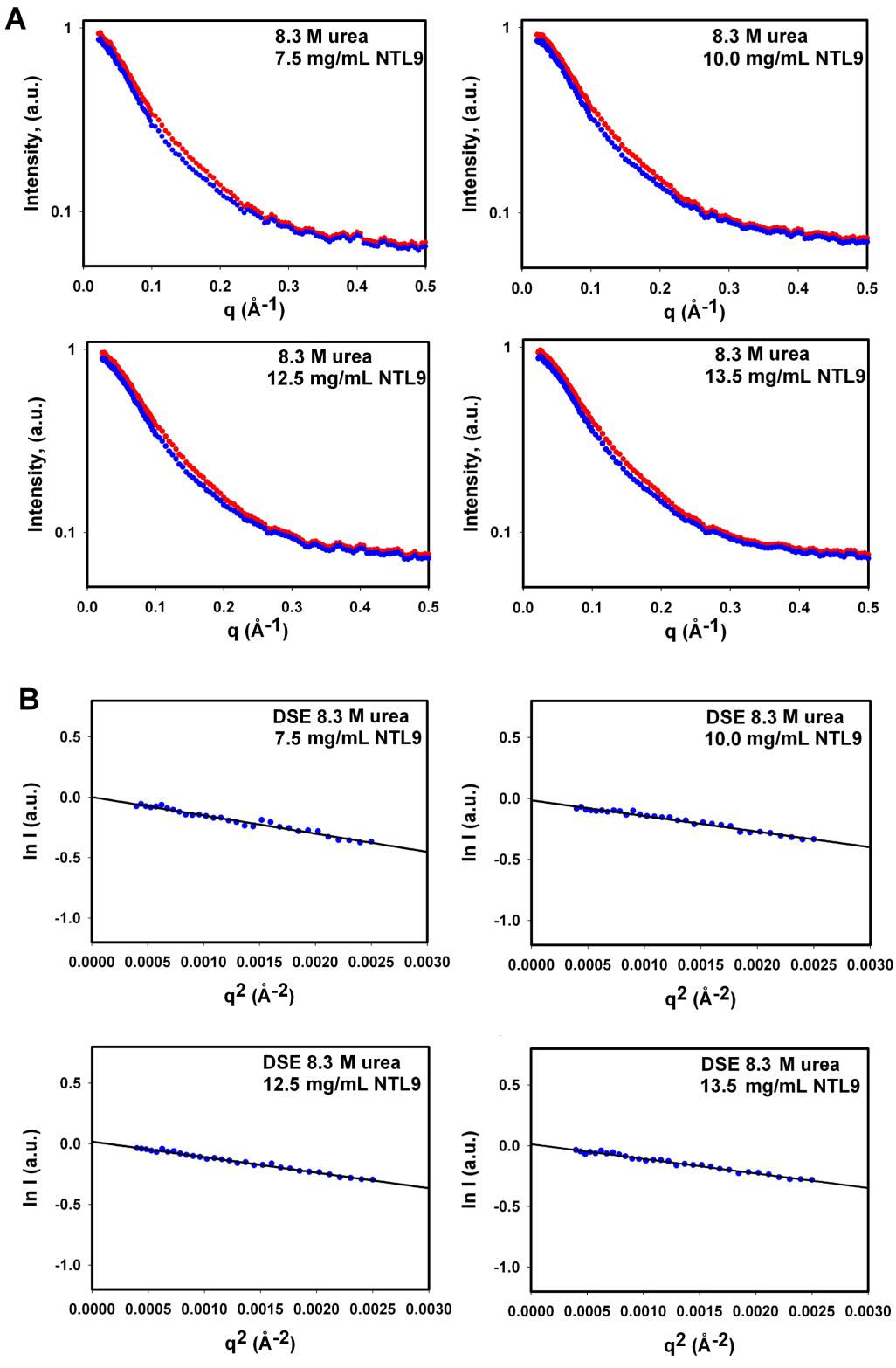

A
8.3 M urea
7.5 mg/mL NTL9
8.3 M urea
10.0 mg/mL NTL9
8.3 M urea
12.5 mg/mL NTL9
8.3 M urea
13.5 mg/mL NTL9
Intensity, (a.u.)
q (Å-1)
B
DSE 8.3 M urea
7.5 mg/mL NTL9
DSE 8.3 M urea
10.0 mg/mL NTL9
DSE 8.3 M urea
12.5 mg/mL NTL9
DSE 8.3 M urea
13.5 mg/mL NTL9
ln I (a.u.)
q2 (Å-2)

Figure 2.1: Determination of R_g for the urea unfolded state of NTL9 in 8.3 M urea at 12°C, pH 5.5. 7% of the protein molecules are folded under these conditions and 93% are unfolded. Experimental scattering profiles were collected for samples of NTL9 at 7.5, 10.0, 12.5, and 13.5 mg/mL in 8.3 M urea and native buffer. The population weighted contribution from the folded state was subtracted from the curves to obtain the scattering profile of the DSE in urea, and these data were used to estimate R_g. (A) Experimental scattering curves collected in urea (red) and curves after subtraction of the native state contribution (blue). (B) Guinier analysis of DSE scattering curves shown in panel A.

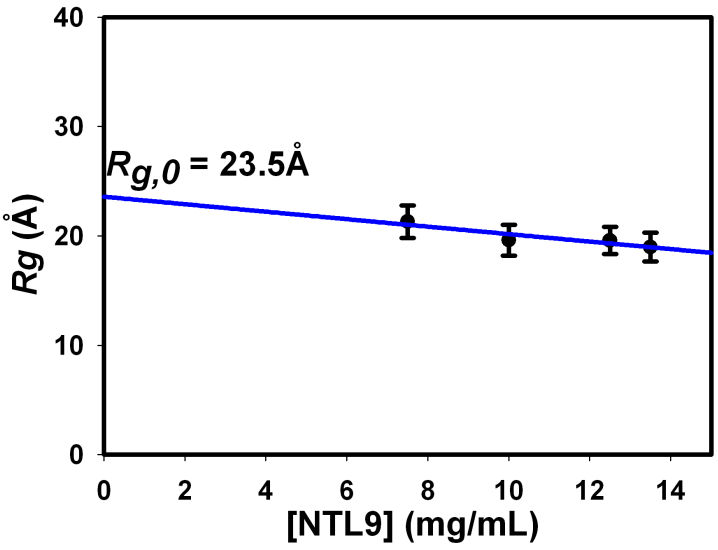


Figure 2.2: Plot of the calculated DSE R_g as a function of protein concentration. The straight line is a linear fit to the data. The value extrapolated to zero concentration is 23.5 +/- 3.5 Å. The uncertainty was estimated by bootstrap analysis.

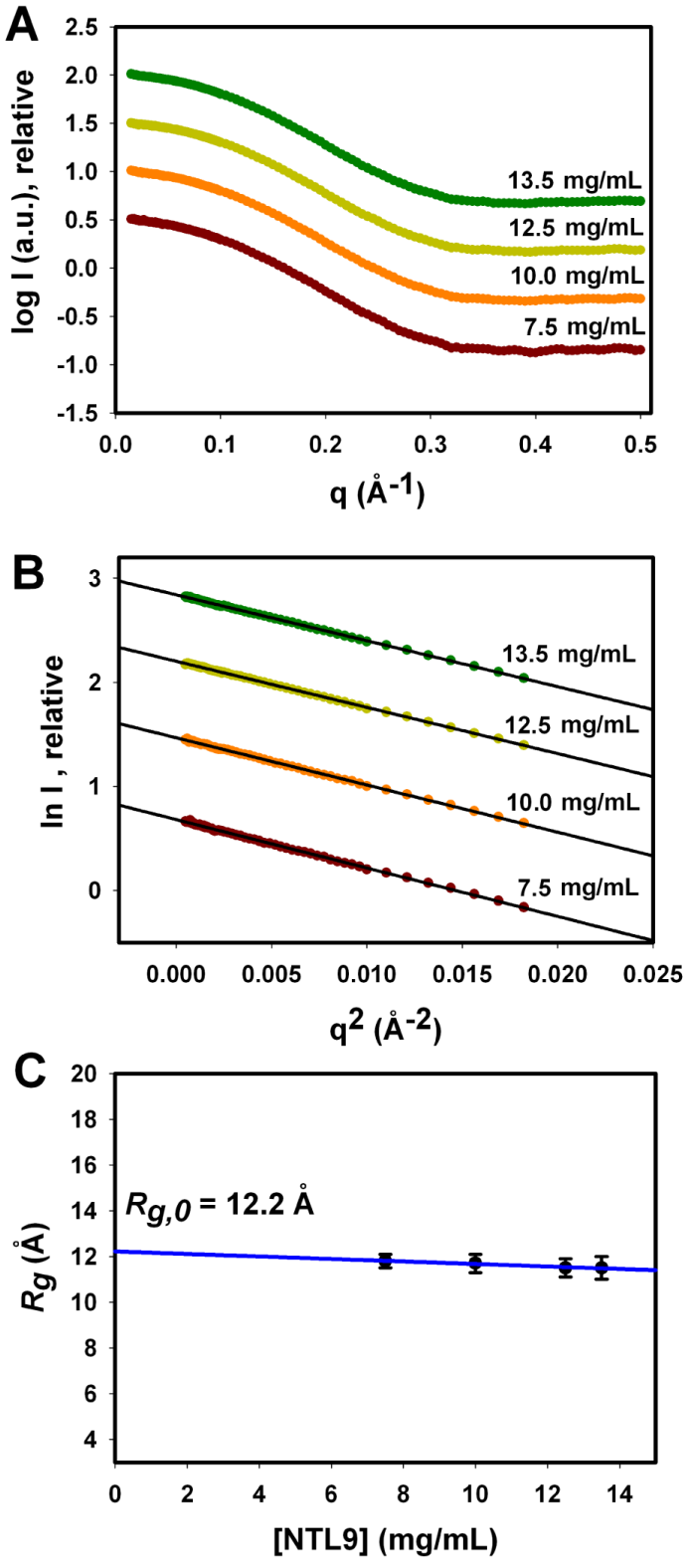

A
log I (a.u.), relative
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0.0
0.1
0.2
0.3
0.4
0.5
q (Å-1)
13.5 mg/mL
12.5 mg/mL
10.0 mg/mL
7.5 mg/mL
B
ln I , relative
3
2
1
0
0.000
0.005
0.010
0.015
0.020
0.025
q2 (Å-2)
13.5 mg/mL
12.5 mg/mL
10.0 mg/mL
7.5 mg/mL
C
Rg (Å)
20
18
16
14
12
10
8
6
4
Rg,0 = 12.2 Å
0
2
4
6
8
10
12
14
[NTL9] (mg/mL)

Figure 2.3: SAXS profile for wild type NTL9 in native buffer at 12°C, pH 5.5. The protein is fully folded under these conditions. The curves are offset from each other for clarity. (A) The scattering curves. (B) Guinier analysis of data shown in panel A. (C) Plot of R_g as a function of protein concentration.

NMR spectra for the native and urea-induced denatured states of NTL9 are shown in Figure 2.4. Although the spectrum is less well resolved in 8.3 M urea, we obtained complete backbone ^{1}H, ^{13}C, and ^{15}N assignments as well as side chain $^{13}C_{\beta}$ assignments using standard approaches. In 8.3 M urea, ninety-three percent of the molecules are unfolded. Although native state resonances were detected at lower contour levels, these were in slow exchange with the denatured state. Secondary shifts [39] were used to quantify the deviation between measured and random coil chemical shifts. Figure 2.5 shows a plot of the $^{13}C_{\alpha}$, $^{13}C_{\beta}$, $\Delta\delta^{13}C_{\alpha} - \Delta\delta^{13}C_{\beta}$, ^{13}CO and $^{1}H_{\alpha}$ secondary shifts for the native state and for the NTL9 DSE in 8.3 M urea. The secondary shifts for the DSE are significantly smaller than those in the native state with all of the ^{13}C secondary shifts being less than 1.0 ppm and the ^{1}H secondary shifts being less than or equal to 0.2 ppm. Residues 42 to 47 exhibit small positive $^{13}C_{\alpha}$ and ^{13}CO secondary shifts secondary shifts (0.31 – 0.37 ppm) as well as small positive values for $\Delta\delta^{13}C^{\alpha} - \Delta\delta^{13}C^{\beta}$ (maximum of 0.34 ppm). This suggests the presence of residual, albeit very small, α-helical propensity. In Figure 2.6 we present a summary of the chemical shifts using secondary structure propensity (SSP) scores [40]. A score of 1 indicates fully formed α-helical structure and a score of –1 represents fully formed β-strands. The SSP scores for the NTL9 DSE in 8.3 M urea are close to zero and support the conclusion of negligible preference for α-helical and β-strand structure.

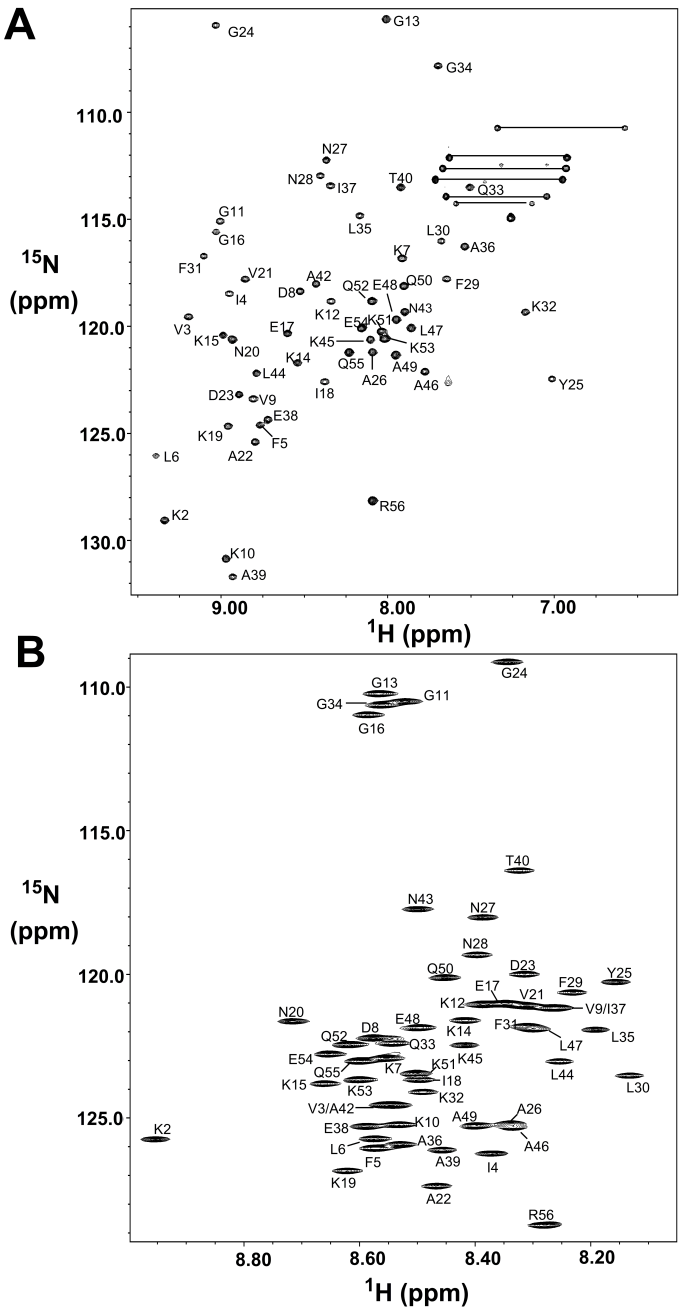


Figure 2.4: HSQC spectrum of NTL9 under folding and denaturing conditions. (A) ^{1}H-^{15}N HSQC spectrum of NTL9 in the fully folded state in 90% H2O/10% D_2O, 20 mM sodium acetate, 100 mM sodium chloride, pH 5.5, 12 ˚C. (B) ^{1}H-^{15}N HSQC spectrum of NTL9 in 8.3 M urea, 20 mM sodium acetate, 100 mM sodium chloride, pH 5.5, 12 ˚C.

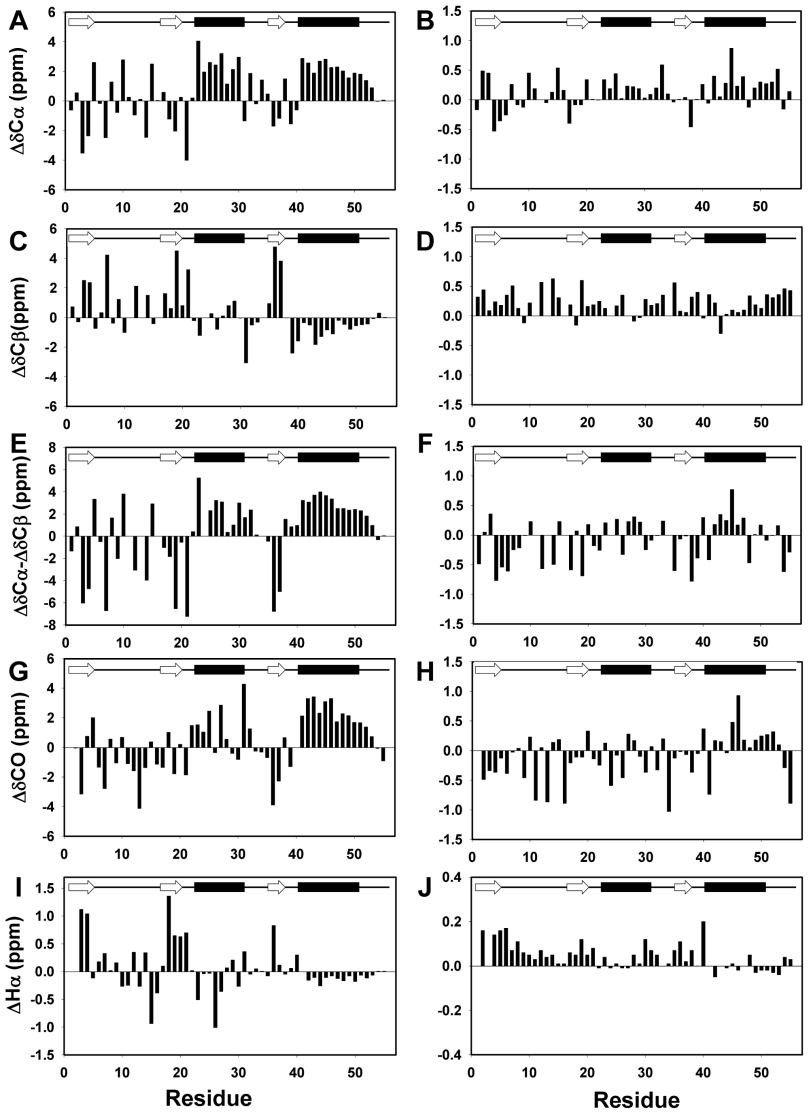


Figure 2.5: Plots of secondary shifts for NTL9 for the fully folded state and for the DSE in 8.3M urea, pH 5.5, 12 °C. Data are plotted as observed minus random coil. (A) Folded $^{13}C_{\alpha}$, (B) DSE $^{13}C_{\alpha}$, (C) Folded $^{13}C_{\beta}$, (D) DSE $^{13}C_{\beta}$, (E) Folded $\Delta\delta^{13}C_{\alpha} - \Delta\delta^{13}C_{\beta}$, (F) DSE $\Delta\delta^{13}C_{\alpha} - \Delta\delta^{13}C_{\beta}$, (G) Folded ^{13}CO, (H) DSE ^{13}CO, (I) Folded $^{1}H_{\alpha}$ and (J) DSE $^{1}H_{\alpha}$. Random coil values of Wishart et al. were used (43). Note the different scales used for the folded and DSE. Schematic diagrams of the elements of secondary structure of the native state are shown at the top of each panel. β-strands are depicted as arrows and α-helices as black bars.

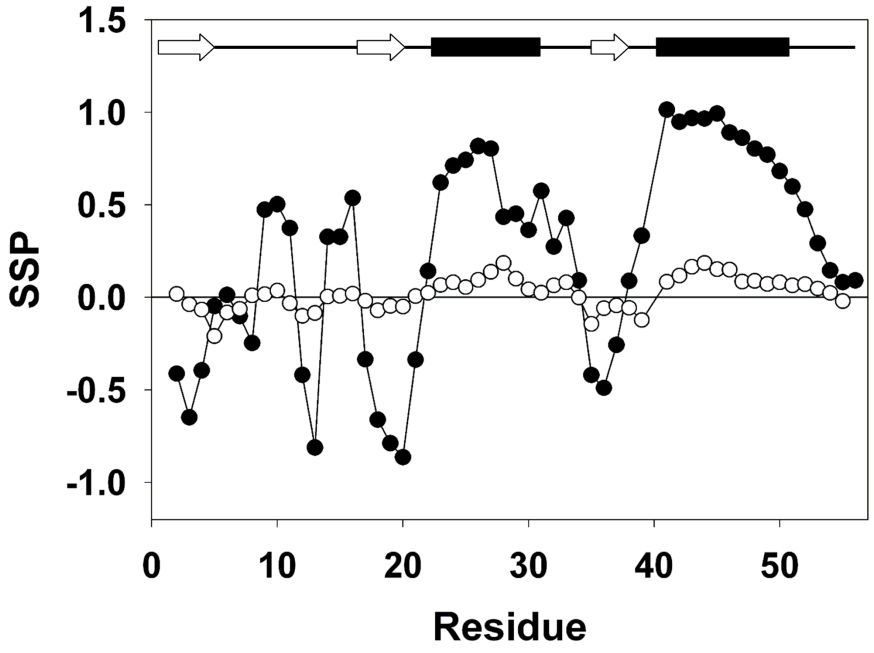


Figure 2.6: SSP scores for the folded (●) and the DSE (○) of NTL9 in 8.3 M urea. A schematic diagram of the elements of secondary structure of the native state is shown at the top of the figure.

2.3 Paramagnetic relaxation enhancement (PRE) experiments indicate the presence of long-range contacts for the NTL9 DSE in 8.3 M urea

Nitroxide spin labels cause significant broadening of NMR resonances of spins that are within 20 Å of the spin label [41-43]. These effects provide a useful probe for long-range contacts [16, 20, 31, 44, 45]. Spin labels were attached at positions 2, 10, 32, 49 and 51, respectively as shown in Figure 2.7. Figure 2.8 shows under native conditions all of the spin-labeled mutants are folded as judged by NMR.

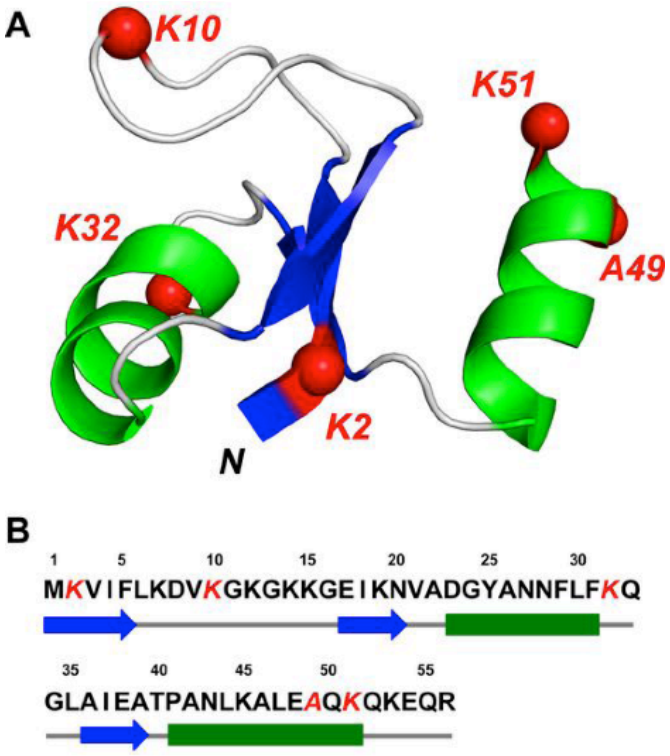


Figure 2.7: Structure of NTL9 with spin label attachment sites. (A) A ribbon diagram of NTL9. The N-terminus is labeled. The points of attachment of the spin labels are indicated as spheres. (B) The primary sequence of NTL9 together with a schematic diagram of the elements of secondary structure where arrows represent β-strands, bars represent α-helices. The sites of attachment of the spin labels are colored in red and are in italics.

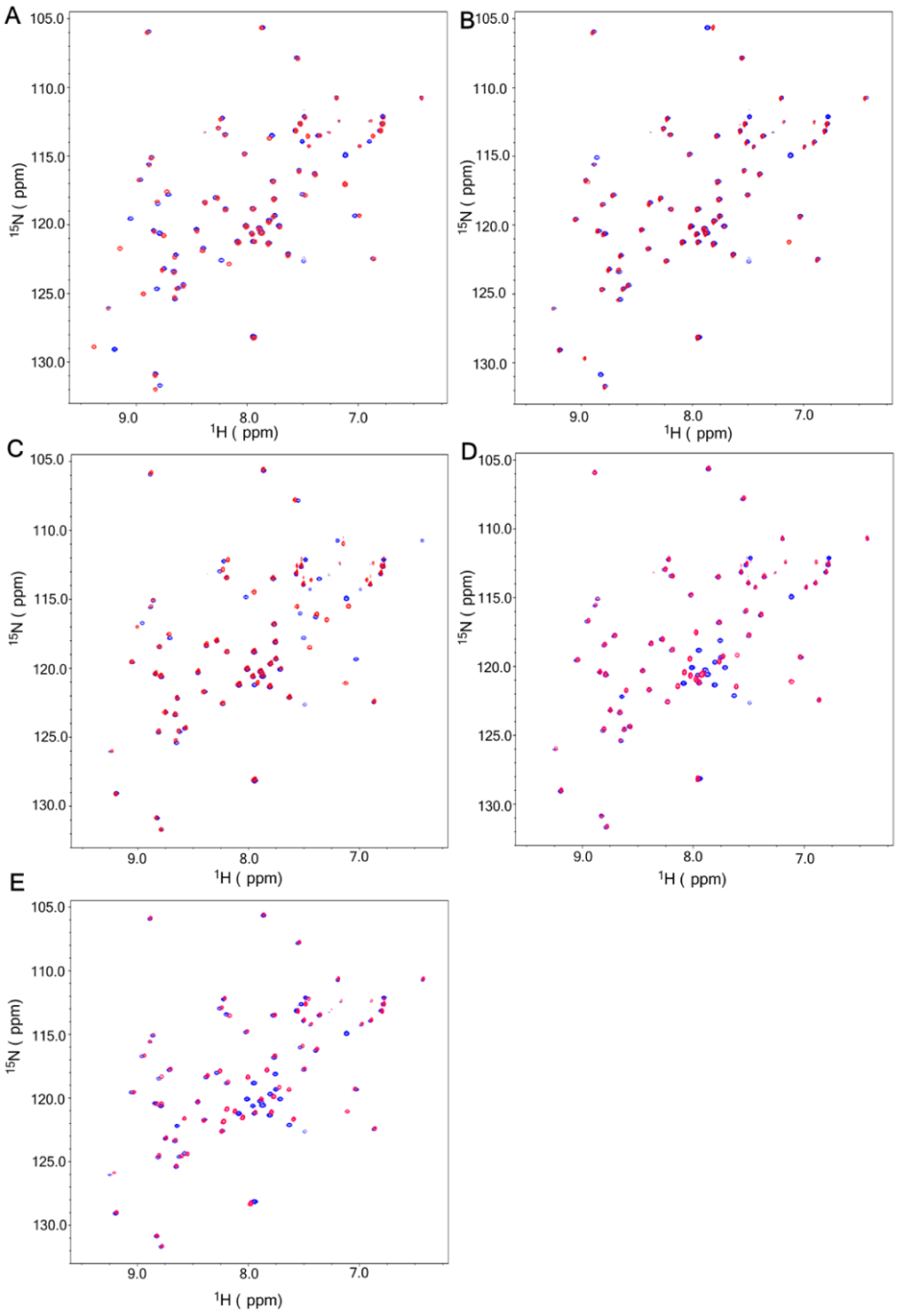


Figure 2.8: Superimposed ^{1}H-^{15}N HSQC spectra of NTL9 wild type (blue) and its cysteine variants (red) for the folded state. (A) K2C, (B) K10C, (C) K32C, (D) A49C and (E) K51C. All NMR spectra were recorded at 12 ˚C, pH 5.5.

We used the peak intensity ratios from the ^{1}H-^{15}N HSQC spectra of the paramagnetic and diamagnetic proteins to quantify PREs. Using the methods of Iwahara et al. [46] we also measured the PRE $^{1}H_N$ transverse relaxation rates (Γ_2) for the NTL9 DSE in 8.3 M urea. Both measurements yield similar trends as shown in Figure 2.9. The PRE results for the DSE show deviations from the values predicted by standard random coil models [41] for all positions. The observed PRE effects are not due to contributions from a small native fraction, since the two states are in slow exchange. We observed similar albeit reduced effects for the variants labeled at residue 49 and 51. There is an asymmetry in the PREs observed for the labels near the N-terminus (K2) compared to the labels near the C-terminus (A49 and K51). We propose that this reflects the differences in contact patterns between the two termini (see simulation results below) and the fact that the spin label is attached to a long side-chain. For example, insertion of the spin labeled side-chain of K2C into a hydrophobic cluster can enhance the relaxation of amide protons in the C-terminus; conversely, the spin labeled side-chains of the A49C and K51C mutants might project away from the clusters leading to decreased relaxation enhancement for amide protons in N-terminus.

The PRE data show significant deviations from the standard Gaussian chain random coil model used to benchmark PREs [41]. The Gaussian random coil model does not account for excluded volume effects nor does it account for the size and flexibility of the attached spin label and it is possible that the observed deviations might reflect limitations of the Gaussian random coil model. To address these issues we performed two sets (with and without spin labels) of atomistic EV limit simulations [17] of NTL9. Results from both simulations differ from the Gaussian chain model and there are small, but detectable differences between the EV simulations with and without spin labels as shown in Figure 2.10. Although the EV limit models predict

more extensive PRE effects than the Gaussian chain model, neither model can reproduce the experimentally observed PREs. The observed PREs cannot be attributed solely to the simplicity of reference models used to calibrate these effects. Instead they suggest the presence of long-range contacts that need quantitative characterization.

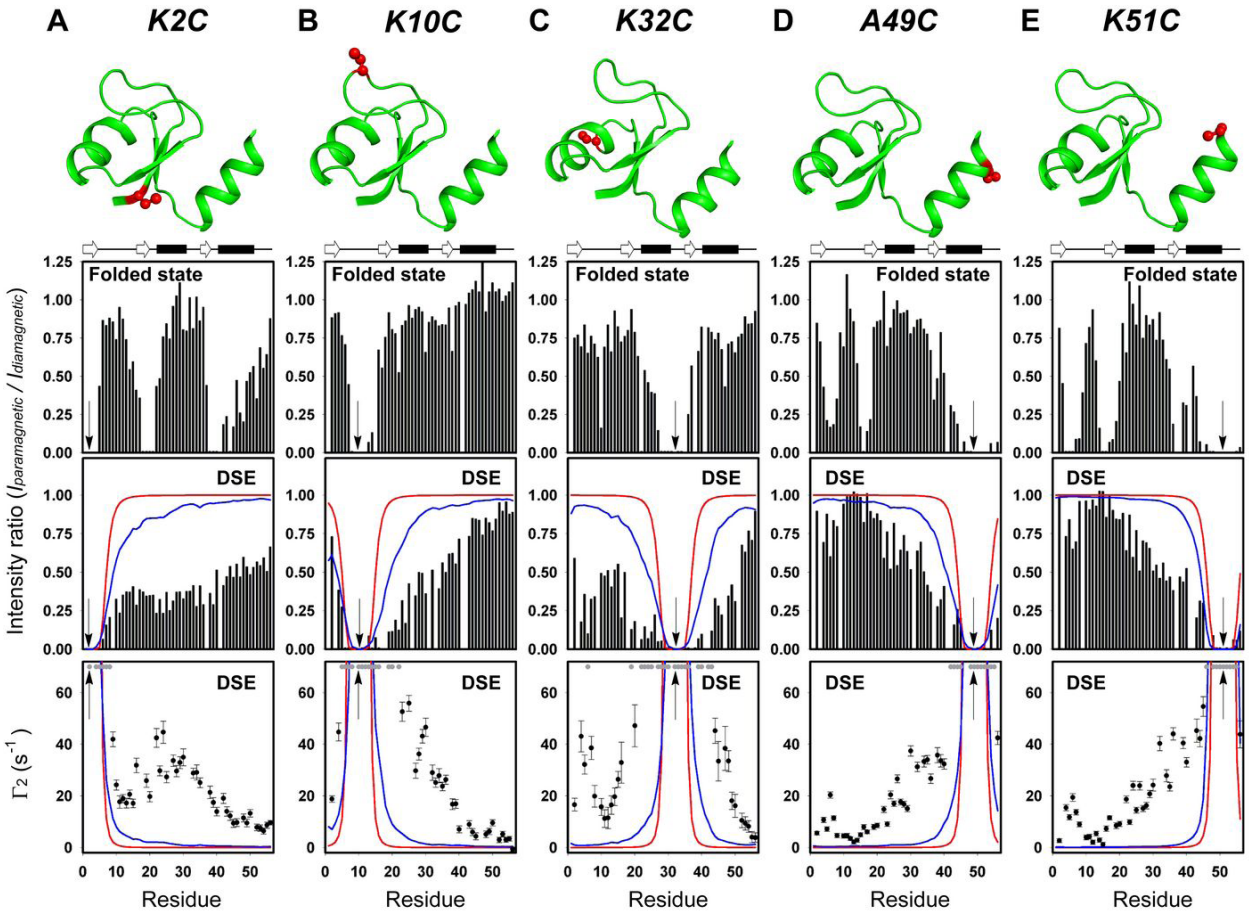


Figure 2.9: Paramagnetic relaxation data for NTL9 under native state conditions and for the DSE in 8.3 M urea. Labeled sites include: (A) K2C, (B) K10C, (C) K32C, (D) A49C and (E) K51C. The histograms display the intensity ratio of the ^{1}H-^{15}N cross-peaks in the HSQC spectra in the folded and urea denatured state. PRE $^{1}H_N$-Γ_2 rates (●) for urea denatured state are plotted in the bottom panels. Residues for which the peaks disappeared in the paramagnetic form are indicated by grey dots. Red lines represent the values expected from the Gaussian random-

coil model. The blue line represents the values calculated using simulations of atomistic ensembles in the EV limit with explicit incorporation of the spin label. An arrow indicates the location of each spin label.

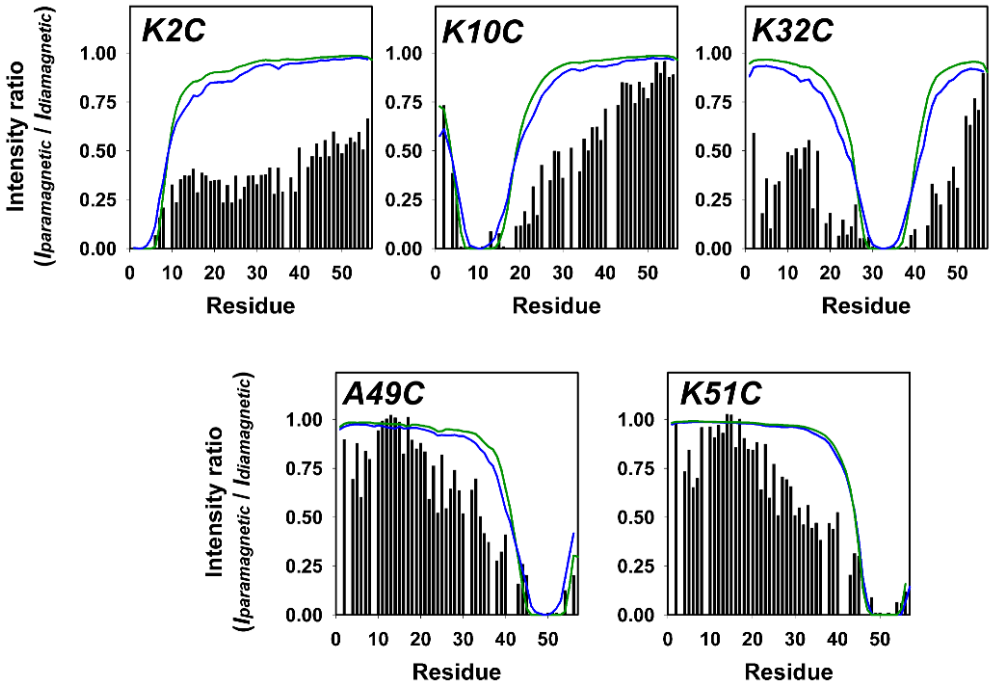


Figure 2.10: Benchmarking models of the DSE. The expected PRE profile was calculated for each of the five spin labeled mutants using an excluded volume model ensemble which explicitly includes the spin label (blue line) and calculated using an excluded volume limit ensemble which does not (green line). For the calculation without the spin label the distances from the C_β carbon of each Cys were calculated to the center of each amide NH bond in the protein backbone. The calculations, which included the spin label, used distances measured from the nitrogen atom on the nitroxide to the center of each amide NH bond in the protein backbone.

2.4 Atomistic simulations help identify ensembles that are consistent with the properties of NTL9 DSEs in 8.3 M urea

We performed Metropolis Monte Carlo (MC) simulations based on atomistic descriptions of NTL9, the ABSINTH implicit solvation model and underlying forcefield paradigm [47], and parameters from the OPLSS-AA/L molecular mechanics forcefield [48]. For each simulation temperature between 240 K and 500 K we performed multiple independent MC simulations, each based on a different random seed and the native state as the starting conformation. We calculated ensemble averages for R_g, secondary structure propensities, and PRE data as a function of the simulation temperature. These were used to identify temperatures whose ensembles generated averages that are concordant with experimental data, and thus serve as models for the DSE of NTL9 in 8.3 M urea.

We compared the measured PRE data to profiles calculated from simulated ensembles using the parameters Δ_1 and Δ_2 that quantify the temperature-dependent root mean square deviations from measured values of the intensity ratios and Γ_2, respectively (see panels A and B in Figure 2.11 and Figure 2.12 for a comparison to intensity ratios). The smallest Δ_j values were obtained for three simulation temperatures 380 K, 390 K and 400 K and we refer to these as T_D temperatures. Figure 2.13 and Figure 2.14 show details of the quantitative agreement between PRE profiles calculated for each of the T_D temperatures and experimental data.

Panel (C) in Figure 2.11 plots the average α-helical and β-sheet contents calculated at each of the simulation temperatures. These contents are less than 3% at T_D temperatures in agreement with estimates from NMR experiments for the DSE of NTL9 in 8.3 M urea. Panel (D) in Figure 2.11 plots the temperature dependence of the ensemble-averaged R_g. We obtained $\langle R_g \rangle$ = 20.50 ± 0.08 Å, 21.21 ± 0.06 Å, and 21.74 ± 0.05 Å for 380 K, 390 K, and 400 K, respectively. These $\langle R_g \rangle$ values are in accord with expectations from scaling relations [10] and similar to the values obtained using SAXS measurements for the DSE of NTL9 in 8.3 M urea. As a reference,

$\langle R_g \rangle$ = 28.31 ± 0.02 Å for NTL9 in the EV limit. We used the ensembles for T_D temperatures as proxies for NTL9 DSE in 8.3 M urea.

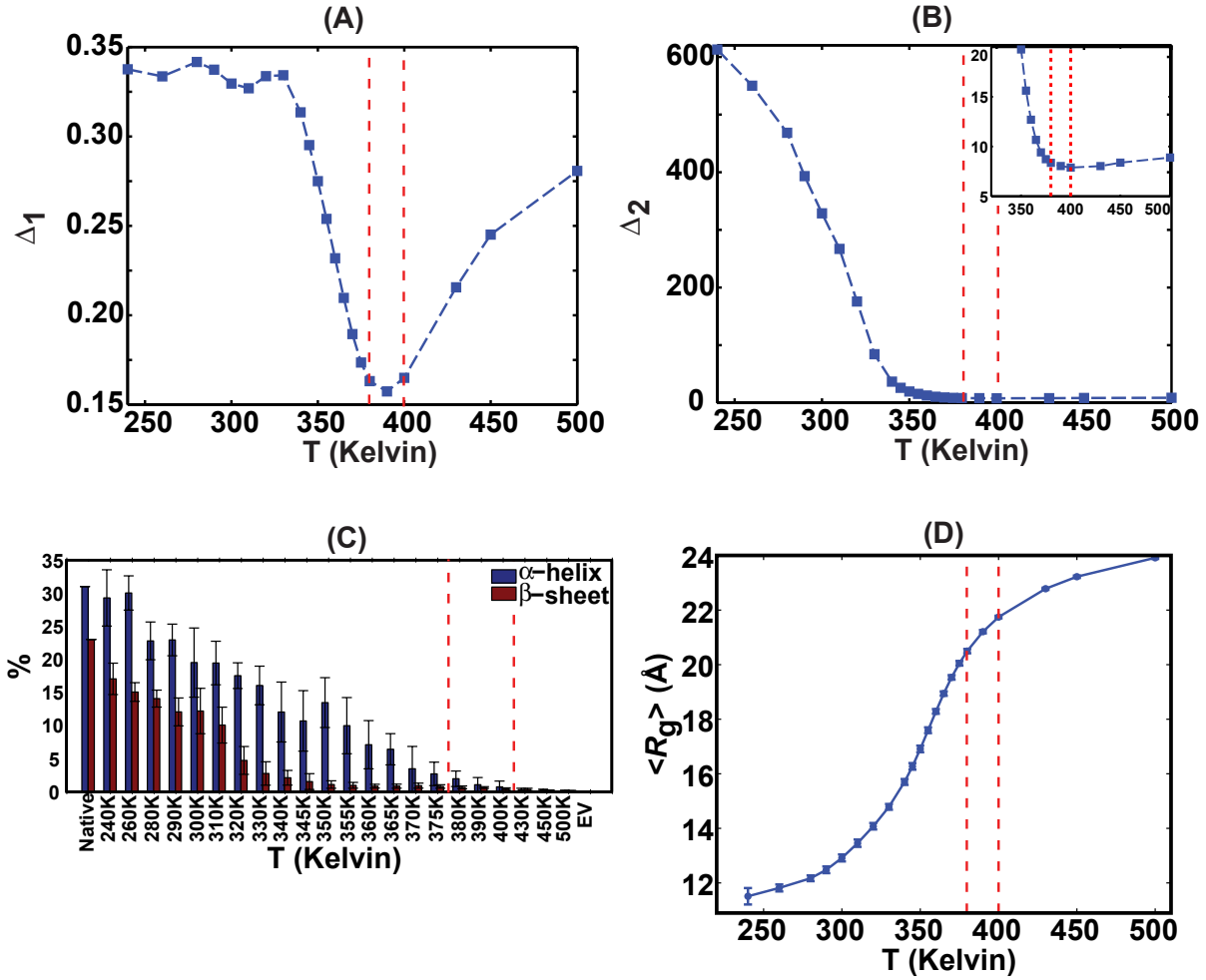


Figure 2.11: Identification of the temperature interval for T_D -ensembles. In all four panels, the dashed lines bracket the temperatures T=380 K, 390 K, and 400 K, which are the T_D temperatures. Panels (A) and (B) show plots of Δ_1 (A) and Δ_2 (B) versus temperature. (C) Plot of the temperature dependence of secondary structure contents, which were calculated using the DSSP algorithm [49]. (D) Plot of the temperature-dependent $\langle R_g \rangle$ values from simulation results.

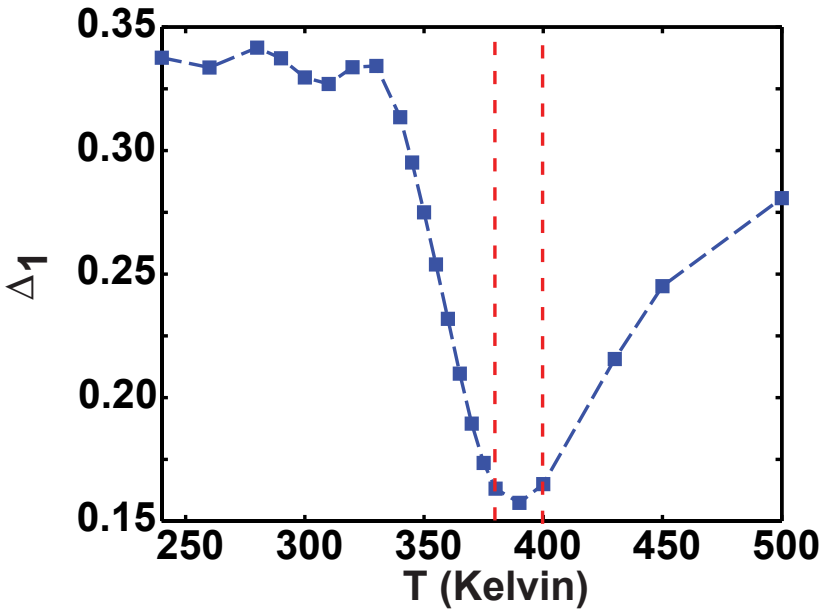


Figure 2.12: Identification of the temperature interval for T_D -ensembles. The dashed lines bracket the temperatures T=380 K, 390 K, and 400 K, which are the T_D temperatures. We compared the measured PRE intensity ratio data to profiles calculated from simulation results using the parameter Δ_1 which is the temperature-dependent root mean square deviation (RMSD) defined in Equation (1) of the main text. The value of Δ_1 is lowest for the temperatures T=380 K, 390 K, and 400 K, respectively.

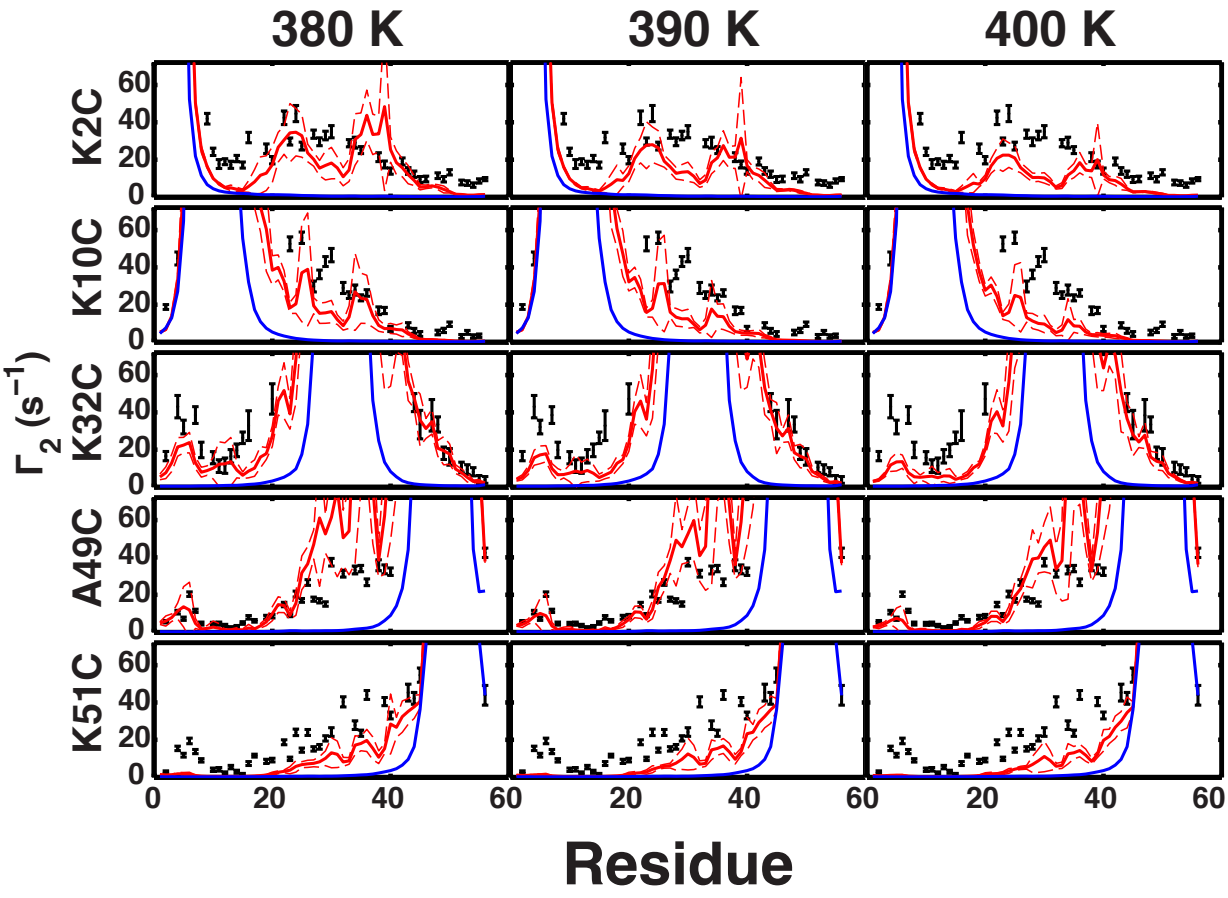


Figure 2.13: Comparison between experimental data (in black symbols with error bars) and calculated PRE profiles, with comparisons shown in terms of Γ_2. The profiles shown in blue were obtained using conformations drawn from the EV ensemble and the profiles shown in red were obtained using conformations for T_D ensembles, with each column corresponding to a specific T_D temperature. The dashed red curves indicate the confidence intervals on calculated profiles. We calculated the latter by partitioning the simulated ensembles into 10 blocks and using the deviations of block averages from the overall mean calculated these errors.

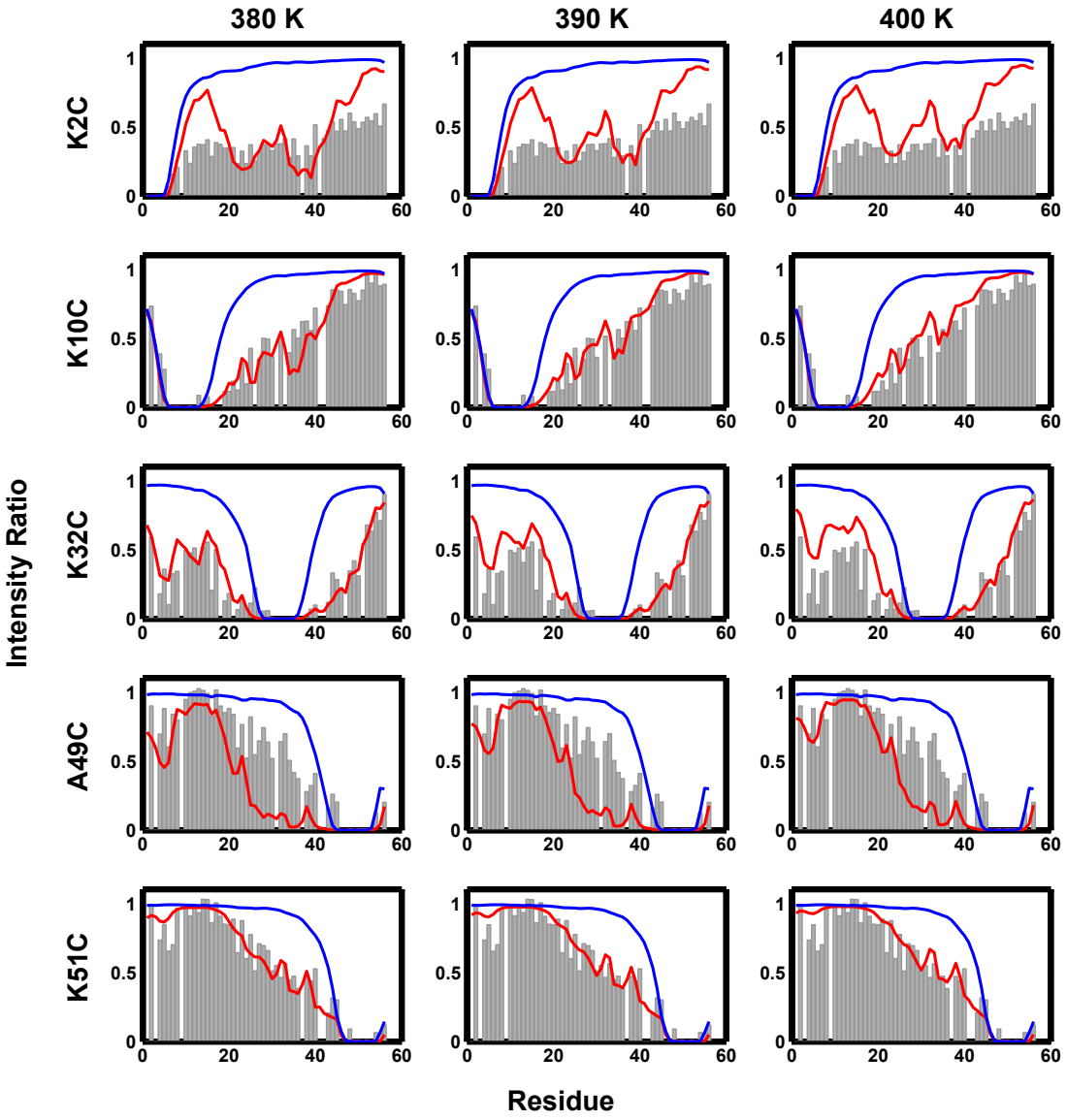


Figure 2.14: Comparison between measured PRE data and calculated profiles. The gray bars denote experimental data for NTL9 in 8.3 M urea – shown here as peak intensity ratios between the diamagnetic and paramagnetic forms. The red curves are the profiles (intensity ratios) calculated using simulated ensembles at each of the T_D temperatures. Each column corresponds to a specific simulation temperature and each row to a specific NTL9 construct that is identified by the sequence position of the spin label in the experiment. In each panel, the blue curves denote the PREs calculated using EV ensembles for NTL9.

2.5 Comparative analysis of contact probabilities

In the EV limit, all pairwise interactions are purely repulsive and as a result the average percent probability p_{ij} of realizing contacts between residues i and j decreases sharply with increasing sequence separation |i–j| such that $0.001\% < p_{ij} < 0.05\%$ for |i–j| > 10 (see Figure 2.15). In contrast, the corresponding probabilities for conformations drawn from the T_D ensembles ($0.01\% < p_{ij} < 1\%$) are at least two orders of magnitude larger than the EV limit for sequence separations in the range $10 < |i–j| < 40$. This indicates the presence of detectable low likelihood medium / long-range contacts in the T_D ensembles.

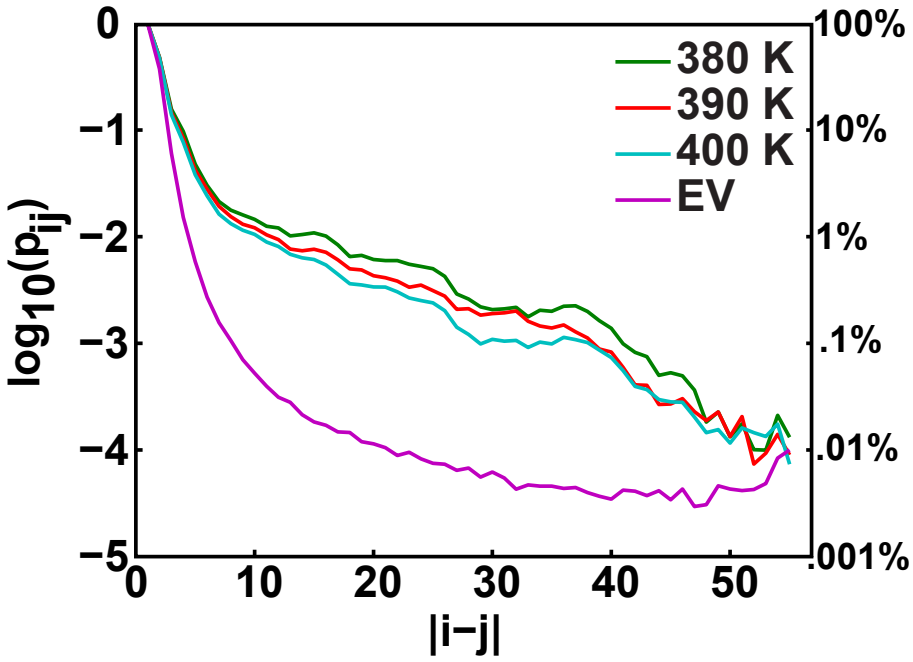


Figure 2.15: Comparison of the probabilities associated with short, intermediate, and long-range contacts. We calculated the probability p_{ij} of realizing spatial contacts between residues i and j that are |i–j| apart in the linear sequence. The figure shows plots of $\log_{10}(p_{ij})$ plotted versus linear sequence separation |i–j| for the three T_D temperatures and the EV ensemble. In all cases, the p_{ij} values decrease with increasing sequence separation |j–i| and the p_{ij} values are less than

1% for |j–i| >10.

2.6 Do T_D temperatures mimic good solvents for NTL9?

The scaling exponent is $\nu \approx 0.59$ in the EV limit and for generic proteins in high concentrations of denaturants [11, 18]. For expanded conformations the average distance $\langle R_{ij} \rangle$ between any pair of residues i and j should follow a power law relationship *viz*., $\langle R_{ij} \rangle = R_o |j-i|^{\nu}$ (Figure 2.16) [11, 17, 21]. If the T_D temperatures mimic good solvents, then ν should approach 0.59 for the scaling of $\langle R_{ij} \rangle$ as a function of |j–i| providing the latter is long enough for scaling theory to apply [17, 21]. For different combinations of pairs of residues (i,j), and (k,l) we used $\nu = \frac{\ln\langle R_{ij} \rangle - \ln\langle R_{kl} \rangle}{\ln(|j-i|) - \ln(|l-k|)}$, where $\langle R_{ij} \rangle$ and $\langle R_{kl} \rangle$ denote average distances and |j–i| and |l–k| denote sequence separations between residues (i,j) and (k,l), respectively. We analyzed the EV limit ensembles of NTL9 for all combinations of (i,j), and (k,l) that satisfy the constraints |j–i| and |k–l| $\geq$ 25. This yields a distribution of values for ν (see panel A of Figure 2.17) with an average value of $\langle \nu \rangle = 0.59 \pm 0.03$. A similar analysis was carried out for each of T_D temperature and we found $\langle \nu \rangle$ to be 0.59 ± 0.05, 0.58 ± 0.05, and 0.57 ± 0.04 for 380 K, 390 K, and 400 K, respectively. Therefore, T_D temperatures mimic good solvent conditions for NTL9.

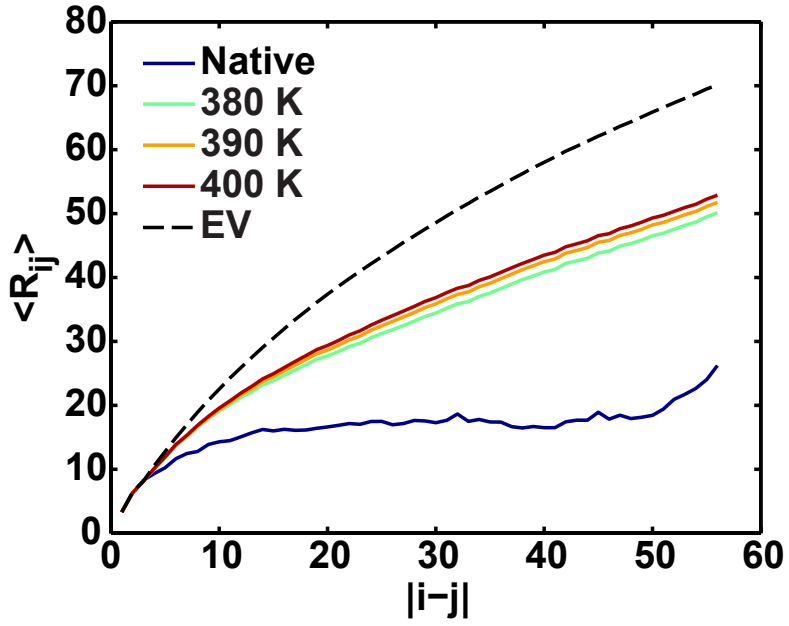


Figure 2.16: Demonstration of the power-law scaling behavior of averaged interresidue distances in the EV limit and T_D ensembles. The average spatial separations between residues i and j show power law dependence on sequence separation |i–j| whereas the folded ensemble (240 K), which is a dominated by compact, globular conformations, shows different, plateauing behavior that is consistent with the non-fractal nature of these ensembles.

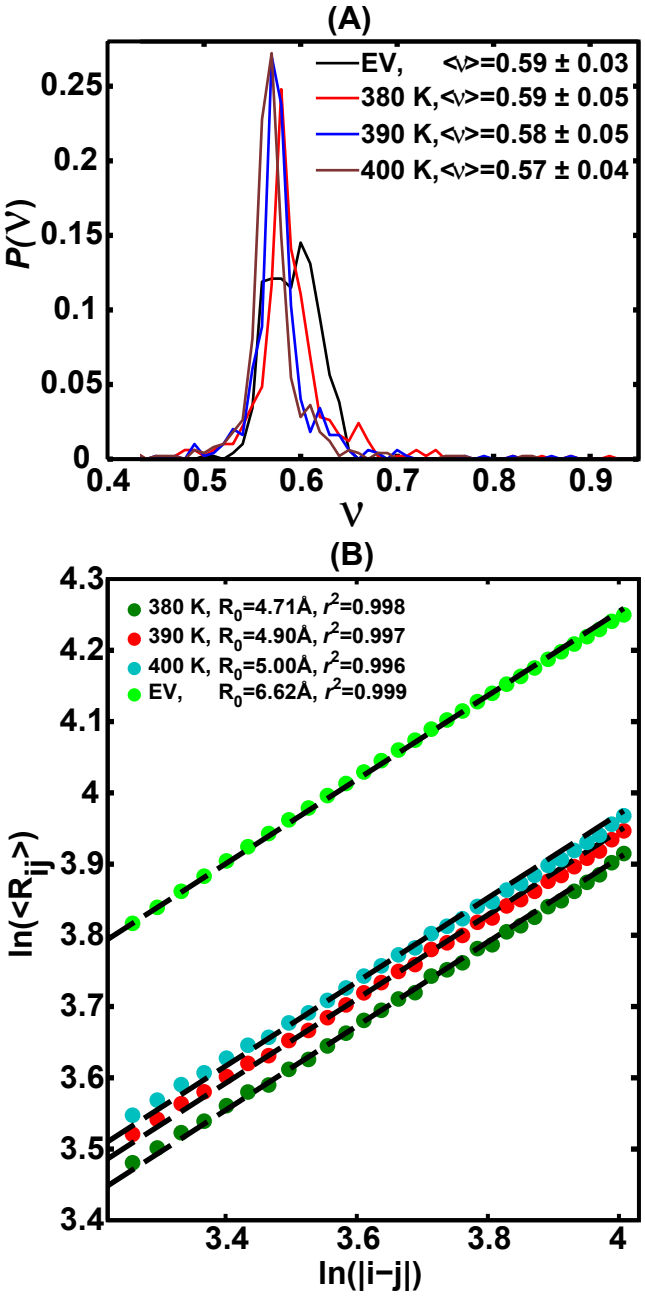


Figure 2.17: Assessing the solvent quality of T_D ensembles. (A) Estimates of the scaling exponent ν. For each combination of i,j and k,l pairs, we obtained a distinct estimate for ν and the analysis over all combinations of pairs leads to a distribution of values, $P(\nu)$, for ν, which are shown here for each of the four T_D temperatures and the EV limit. In all of the cases the bin

widths are 0.01. The constraints on |j–i| and |l–k| were chosen by requiring $P(\nu)$ to be a Gaussian distribution for the EV limit as assessed by the D'Agostino-Pearson omnibus test [50] and a *p*-value of 0.05. (B) Estimates of R_o using linear regression to analyze results for the scaling of internal distances with sequence separation by fitting these results to the equation $\ln(\langle R_{ij}\rangle) = \ln(R_o) + 0.59\ln(|j-i|)$. The circles are simulation results for $\ln(\langle R_{ij}\rangle)$ and dashed lines are fits to the data. The legend shows the values obtained for R_o and the coefficient that measures the correlation between each linear fit and the corresponding data set. Our estimates for R_o will be larger than the intercept for the scaling of $\langle R_g\rangle$ – it should be approximately $\sqrt{1/6}$ times the intercept for the scaling of internal distances. Multiplying our R_o values by $\sqrt{1/6}$ yields values for $\langle R_g\rangle$ that are similar to the estimate of Kohn et al. [10].

2.7 Reconciling $\nu \approx 0.59$ with deviations from the EV limit

Dimensionless quantities such as $|j-i|^{0.59}$ are converted to distances using a multiplicative pre-factor R_o. In a good solvent, this parameter can be used to quantify the average volume excluded per residue for favorable interactions with the surrounding solvent. If the net charge per residue is ≤ 0.3 [19] and intra-chain electrostatic repulsions are screened [51], then $R_o \leq R_o^{EV}$ in a generic good solvent where R_o^{EV} denotes the value of R_o calculated in the EV limit. If $R_o < R_o^{EV}$, then residual intra-chain attractions persist in the good solvent. This will influence the probabilities associated with long-range contacts.

We estimated the value of R_o for each T_D temperature and compared these to R_o^{EV}. We fit the values for $\langle R_{ij}\rangle$ to the equation $\ln(\langle R_{ij}\rangle) = \ln(R_o) + 0.59\ln(|j-i|)$ for all $|j-i| \geq 25$ where R_o is the free parameter. The results are shown in panel (B) of Figure 2.17. The maximal value of R_o is

R_o^{EV} and these values decrease as the T_D temperatures decrease. Our analysis provides quantitative evidence for imperfect compensation between non-EV-intrachain and chain-solvent interactions. For the DSE of NTL9 our model ensembles suggest that imperfect compensation results from the presence of residual intrachain attractions.

2.8 Analysis of contact patterns for T_D temperature ensembles

Figure 2.18 quantifies the probabilities of inter-residue contacts in the native state ensemble (240 K), the T_D temperatures, and the EV limit, respectively. Each cell in a contact map quantifies the probability that residues i and j are in contact. We define residues to be in contact if the inter-residue distance between at least one pair of heavy atoms is ≤ 3.5Å. We observe patterns that include several low-probability native as well as non-native contacts (p_{ij} < 0.1). We calculated two sets of difference contact maps to quantify the differences between the ensembles sampled at T_D temperatures and the native state and EV limit ensembles, respectively. Figure 2.19 and Figure 2.20 show difference maps with respect to the native state ensemble and the EV limit as the reference. The colors of the cells are set by the magnitude and sign of the quantity $d_{ij} = p_{ij} - q_{ij}$, where q_{ij} denotes the probability that residues i and j are in contact in the reference ensemble (native or EV). If $d_{ij} < 0$, then the probability of realizing a contact between residues i and j is lower at the T_D temperatures when compared to the reference ensembles and the converse is true if $d_{ij} > 0$.

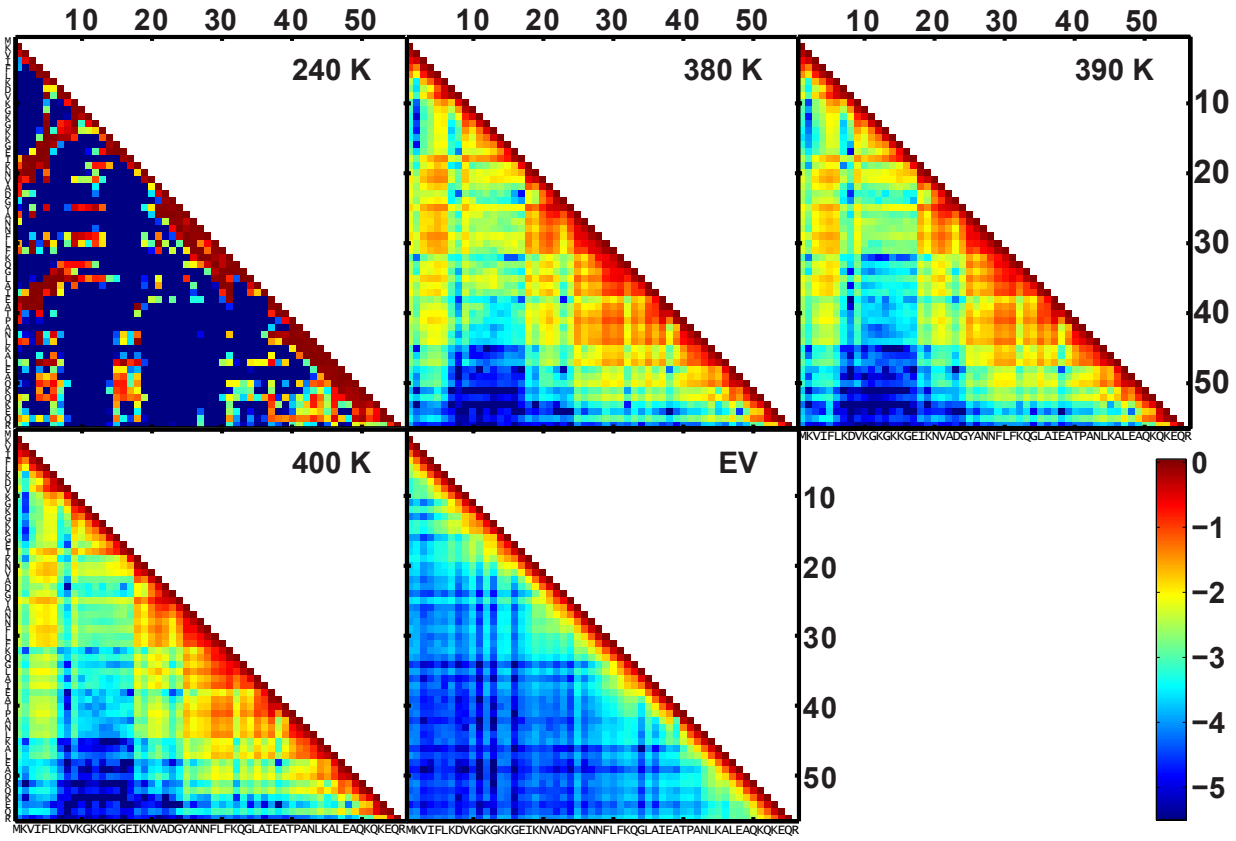


Figure 2.18: Contact maps for the native state ensemble (240 K), the T_D temperatures, and the EV limit, respectively. In the contact maps, the coloring in each cell denotes the probability, p_{ij}, of realizing a contact between residues i and j. Only the lower triangular maps are shown because the contact maps are symmetric. To facilitate a clear quantitative comparison between different ensembles, the plots show $\log_{10}(p_{ij})$ instead of p_{ij}. A similar color bar is used for each contact map and this is shown in the bottom right corner.

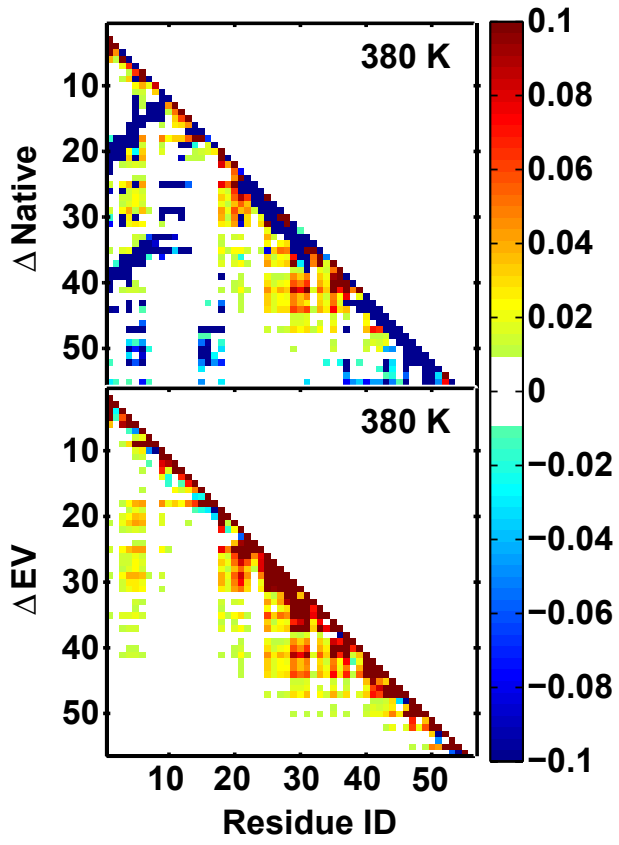


Figure 2.19: Difference contact maps for 380 K. The top row shows difference contact maps with respect to the native state ensemble (240 K) and the bottom row shows similar maps with respect to the EV ensemble. The cooler colors imply that the contact in question has a higher probability in the reference ensemble (native or EV) whereas the warmer colors imply that the contact has a higher probability at 380 K when compared to the reference ensemble. A value of zero implies that the contacts either have similar probabilities in both the T_D and references ensembles or are missing in both.

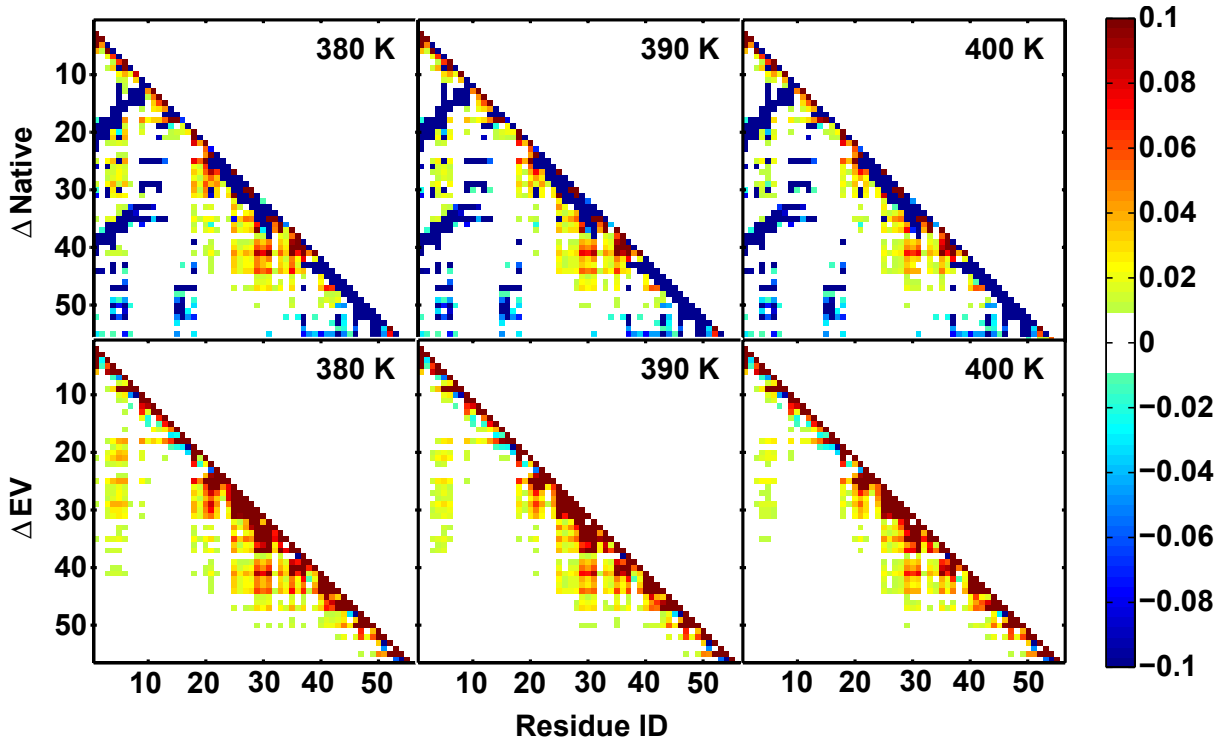


Figure 2.20: Difference contact maps for each T_D temperature. The top row shows the difference contact maps with respect to the native state ensemble (240 K) and the bottom row shows similar maps with respect to the EV ensemble. For each T_D temperature the difference contact map is obtained by calculating $d_{ij} = p_{ij} - q_{ij}$, where p_{ij} is the average contact probability for a pair of residues i and j at a specific T_D temperature and q_{ij} is the probability for the corresponding contact in the reference ensemble (native or EV). In the difference maps, the cooler colors imply that the contact in question has a higher probability in the reference ensemble (native or EV) whereas the warmer colors imply that the contact has a higher probability in ensembles corresponding to the T_D temperatures when compared to the reference ensemble. A value of zero implies that the contacts either have similar probabilities in both the T_D and references ensembles or are missing in both.

The prominent differences between T_D ensembles and the two reference ensembles are the higher probabilities associated with long- and intermediate-range contacts involving

hydrophobic residues drawn from four specific groups labeled $g_1 - g_4$ where $g_1 \equiv$ (M1, V3, I4, F5, L6), $g_2 \equiv$ (I18, K19, N20, V21, A22), $g_3 \equiv$(G24, Y25, A26, N27, N28, F29, L30, F31), and $g_4 \equiv$(G34, L35, A36, I37). These contacts have low probabilities (≤ 0.1), are predominantly non-native (see top row in Figure 2.19 and Figure 2.20), and are either medium-range (contacts between residues from g_1 and g_2, g_2 and g_3 or g_3 and g_4) or long-range (contacts between residues from g_1 and g_3 or g_1 and g_4). Figure 2.21 displays a montage of conformations to illustrate how medium / long-range, low-probability contacts can form without requiring either persistent secondary structure or global compaction at T_D temperatures.

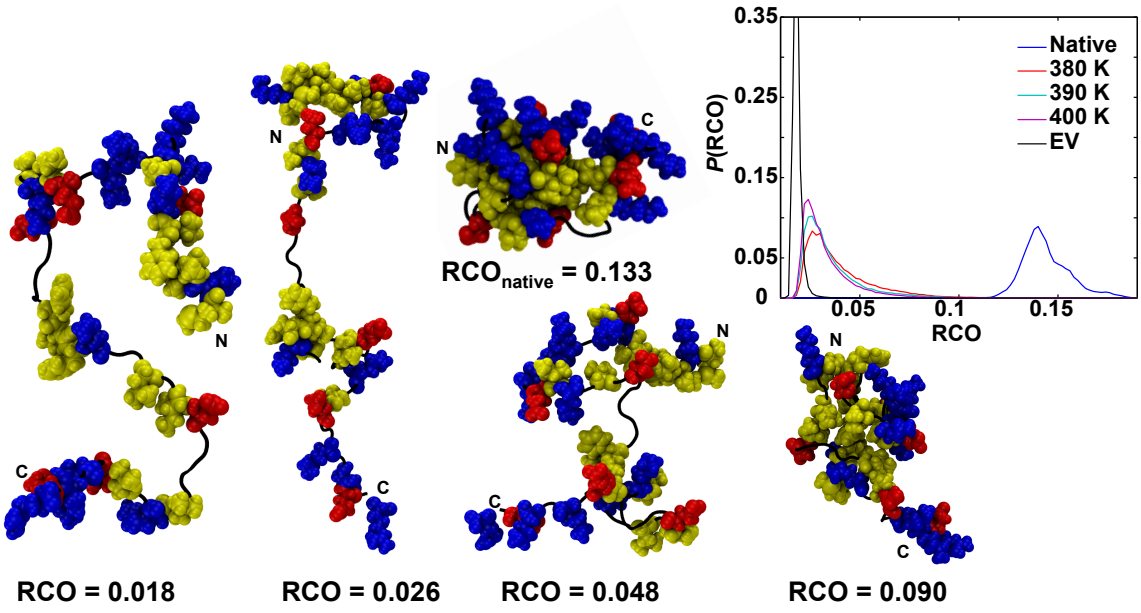


Figure 2.21: Distributions (top right panel) of the relative contact order (RCO) for different ensembles and representative conformations drawn from the RCO distribution at 390 K. $\text{RCO}=\frac{1}{N}\sum_{i=1}^{n_{\text{contacts}}} s_i$; Here, N is the number of residues, n_{contacts} is the number of contacts in a specific conformation, and s_i is the sequence separation between the pair of residues that make up contact i. The bin widths are 0.002 RCO units, which are dimensionless. The overlap

fractions between distributions for T_D temperatures and the EV and native state ensembles are as follows: (0.0935, 0.1136, 0.1325) and (0.0032, 0.0036, 8.33×10^{-4}, 1.83×10^{-4}) for 380 K, 390 K, and 400 K, respectively. In each snapshot, the polypeptide backbone is shown using a black contour. Positively charged sidechains are shown in blue, negatively charged sidechains in red, and hydrophobic groups in yellow. Each conformation drawn from the T=390 K ensemble is annotated by its corresponding RCO value *viz*., 0.018, 0.026, 0.048, and 0.090. The N- and C-termini are labeled in each snapshot. In addition, the figure also includes a snapshot drawn from the T=240 K, native state ensemble with an RCO value of 0.133.

2.9 Discussion

We have shown that imperfect compensation between non-EV-intrachain and chain-solvent interactions is compatible with $N^{0.59}$ scaling for denatured proteins. This is manifest as detectable, low-probability ($p_{ij} < 0.1$), medium / long-range contacts which our analysis ascribes to residual intra-chain attractions between specific clusters of hydrophobic residues. Interestingly, the residues involved in many of the low probability non-native medium / long-range contacts are highly conserved among different NTL9 sequences (Figure 2.22). The presence of such contacts may therefore be a conserved feature among members of the NTL9 family. Furthermore, the $\langle R_g \rangle$ values for proteins in the EV limit [17, 18] are larger than those obtained for denatured proteins using SAXS [10]. This implies the presence of residual intra-chain attractions in most denatured proteins. Indeed, Wu et al. [3] have proposed that clusters of Ile, Leu, Val, and Phe are present in unfolded states because they form early during folding. Weak intra-chain attractions can reshape the DSE vis-à-vis the EV limit and do so without altering the $N^{0.59}$ scaling behavior. These interactions might reduce the likelihoods associated with deleterious intermolecular interactions between unfolded proteins that lead to protein aggregation

[5], modulate internal friction in the denatured state [52] and reshape barriers to protein folding [53].

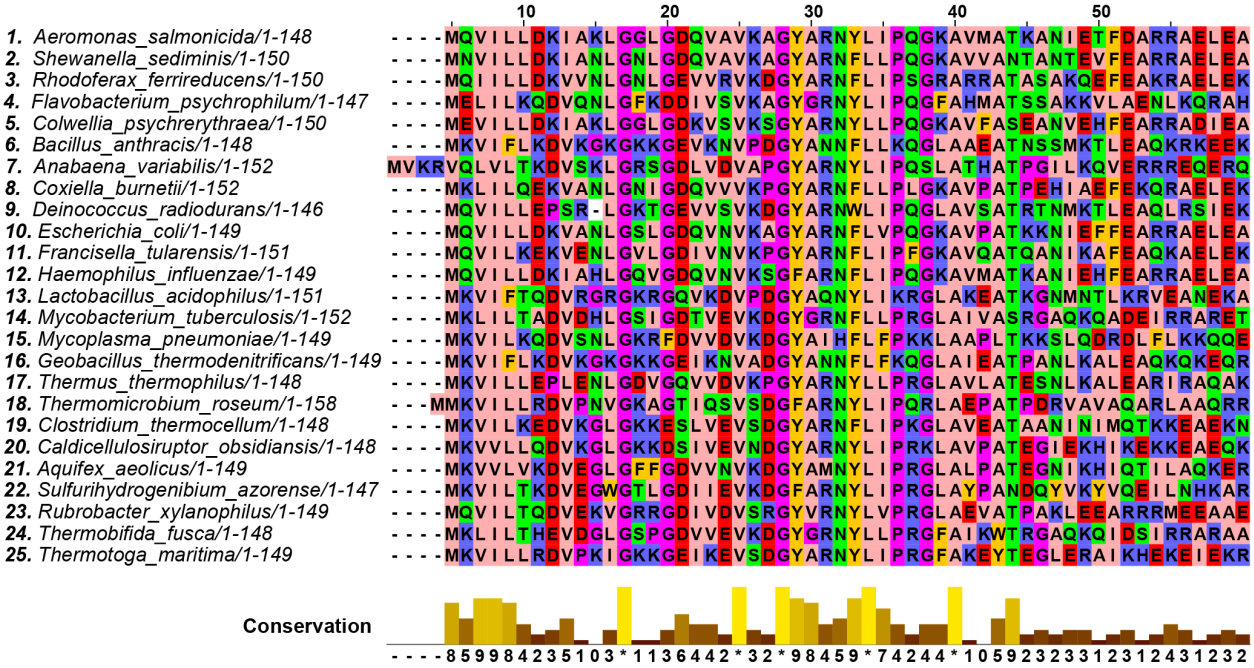


Figure 2.22: Alignment of bacterial N-terminal NTL9 sequences. Sequences 1-5 are psychrophiles, 6-15 are mesophiles, and 16-25 are thermophiles. Hydrophobic clusters and their relative positions are conserved for all bacterial strains suggesting a functional role.

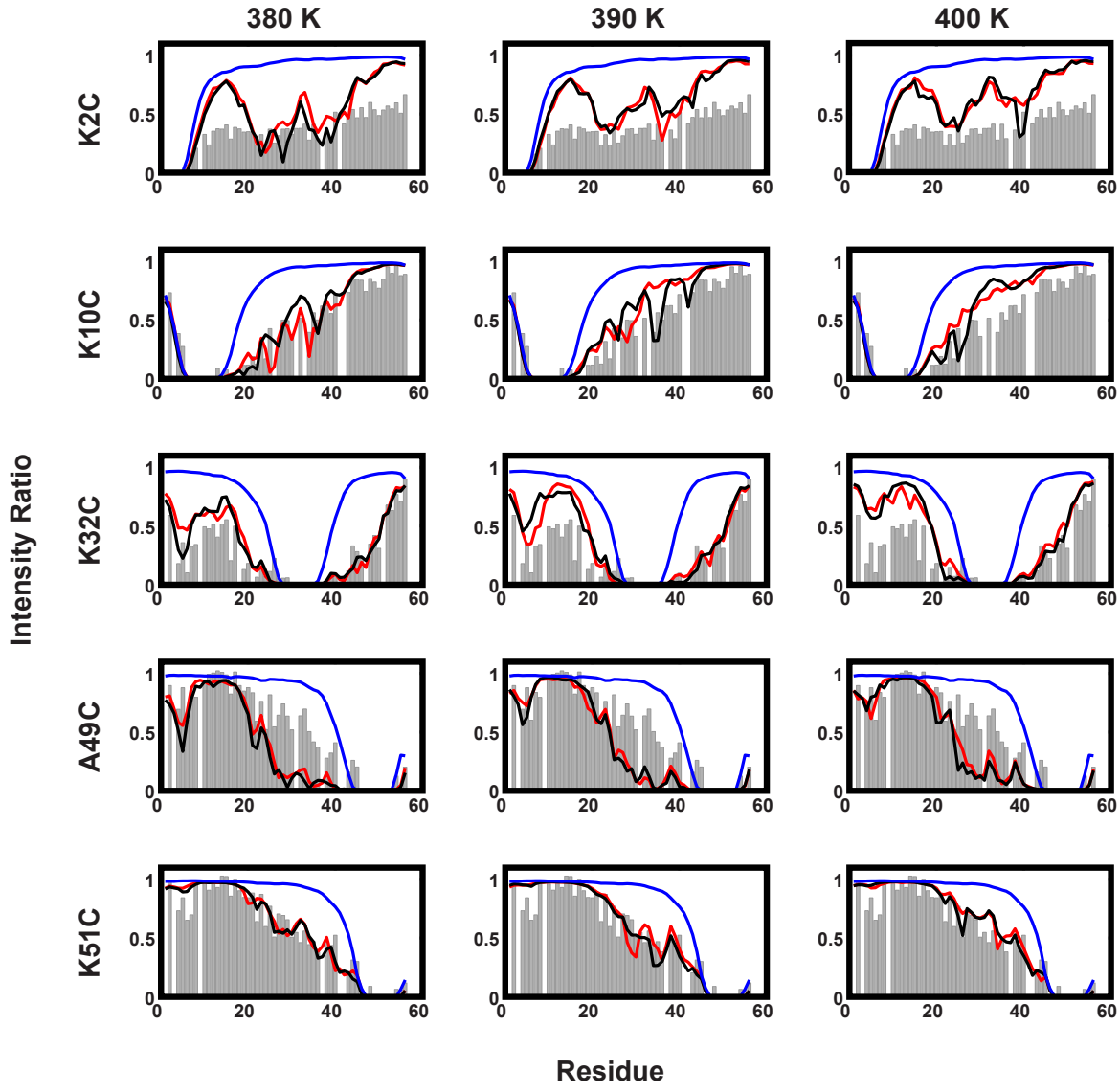


Figure 2.23: Quantification of the dependence of calculated PRE profiles on salt (NaCl) concentration. The comparisons are shown in terms of the intensity ratios. The calculated profiles are as follows: EV limit (blue), 25 mM (red), and 120 mM (black).

2.10 Conclusion

The sizes of unfolded proteins under highly denaturing conditions scale as $N^{0.59}$ with chain length. This suggests that denaturing conditions mimic good solvents whereby the preference for favorable chain-solvent interactions causes intra-chain interactions to be repulsive,

on average. Beyond this generic inference, the broader implications of $N^{0.59}$ scaling for quantitative descriptions of denatured state ensembles (DSEs) remain unresolved. Of particular interest is the degree to which $N^{0.59}$ scaling can simultaneously accommodate intra-chain attractions and detectable long-range contacts. Here, we present data showing that the DSE of the N-terminal domain of the L9 ribosomal protein in 8.3 M urea lacks detectable secondary structure and forms expanded conformations in accord with the expected $N^{0.59}$ scaling behavior. Paramagnetic relaxation enhancements, however, indicate the presence of detectable long-range contacts in the denatured-state ensemble of NTL9. To explain these observations we used atomistic thermal unfolding simulations to identify ensembles whose properties are consistent with all of the experimental observations thus serving as useful proxies for the DSE of NTL9 in 8.3 M urea. Analysis of these ensembles shows that residual attractions are present under mimics of good solvent conditions and for NTL9 they result from low likelihood, medium/long-range contacts between hydrophobic residues. Our analysis provides a quantitative framework for the simultaneous observation of $N^{0.59}$ scaling and low likelihood long-range contacts for the DSE of NTL9. We propose that such low likelihood intramolecular hydrophobic clusters might be a generic feature of DSEs that play a gatekeeping role to protect against aggregation during protein folding.

2.11 Materials and methods

2.11.1 Protein expression and purification

^{15}N-labeled and ^{13}C,^{15}N-labeled NTL9 wild type and mutants were expressed in *Escherichia coli* BL21 cells in M9 minimal medium using standard methods. For ^{15}N-labeled proteins 0.8 g/L $^{15}NH_4Cl$ was used as the sole nitrogen source and for ^{13}C, ^{15}N-labeled protein 4 g/L [^{13}C]glucose was used as sole carbon source. The cells were grown at 37 ˚C until the OD at

600 nm reached 0.8, and then induced with 1 mM IPTG for 4 h. Cells were harvested and lysed by sonication. The proteins were purified with an ion-exchange column, and then by reverse-phase HPLC. The identity of each protein was confirmed by mass spectrometry.

2.11.2 Small angle X-ray scattering (SAXS) experiments

Samples of NTL9 were prepared in native buffer (20 mM sodium acetate and 100 mM sodium chloride, pH 5.5), as well as in 8.3 M urea. Scattering experiments were performed at beamline X9 at Brookhaven National Laboratory, National Synchrotron Light Source I (Upton, New York, USA). Protein samples were injected into a 1 mm diameter capillary continuously during the measurement at a rate of 0.67 μL/s in order to avoid radiation damage. The exposure time for each measurement was 30 s. Scattering data was collected for four separate protein concentrations: 7.5, 10.0, 12.5, and 13.5 mg/mL at 12 ˚C. Each sample was measured three times and then averaged before data analysis. The program PRIMUS [54] was used for buffer subtraction, and the radius of gyration (R_g) was obtained using the Guinier approximation, Equation 2.1,

$$I(q) = I(0)\exp(-R_g^2 q^{2/3}) \tag{2.1}$$

where $I(q)$ is the intensity at scattering vector q [55].

2.11.3 NMR sample preparation

Samples for native state studies were prepared in 90% H_2O/10% D_2O with 20 mM sodium acetate and 100 mM sodium chloride, pH 5.5. For denatured state studies urea was added to a final concentration of 8.3 M determined by refractometry. All NMR experiments were recorded at 12 ˚C, and protein concentration of approximately 1 mM. 2,2-Dimethyl-2-silapentane-5-sulfonate sodium salt (DSS) was added as an internal reference (0.00 ppm) for all samples.

2.11.4 NMR assignments

The ^{1}H offset frequency was centered at the water resonance and the ^{15}N offset frequency was set at 118.0 ppm. 2D ^{1}H-^{15}N HSQC, 3D HNCACB, CBCA(CO)NH, HNCO, and ^{1}H-^{15}N TOCSY-HSQC experiments were performed to generate assignments for the native state. 2D ^{1}H-^{15}N HSQC and 3D HNCACB, CBCA(CO)NH, HNCO, HNCA and ^{1}H-^{15}N TOCSY-HSQC were used to obtain the DSE assignments. Data was processed with the NMRPipe package [(56)], and chemical shift assignments were accomplished using NMRViewJ [(57)]. ^{1}H chemical shifts were referenced to DSS directly, and ^{15}N, ^{13}C chemical shifts were indirectly referenced using standard methods. The random coil values of Wishart and coworkers were used to calculated secondary chemical shifts [(58)]. Sequence-dependent correction of random coil chemical shifts were performed [(59)].

2.11.5 Pulsed-field gradient NMR diffusion experiments

A Bruker 600 MHz spectrometer equipped with a cryoprobe was used for the Pulsed-field Gradient NMR Diffusion experiments. Protein samples dissolved in deuterated urea were exchanged in 100% D_2O for 5 h at 25 ˚C and lyophilized. The exchange procedure was repeated three times to ensure that amide protons in the protein and the protons in urea were fully exchanged. The final concentration of urea was 8.3 M as measured by refractometry and the pD was adjust to 5.1 (uncorrected pH meter reading). 1, 4-dioxane was used as an internal standard. A pseudo-2D version of PFG- diffusion-ordered spectroscopy (DOSY), was used for data collection and analysis. The diffusion delay, Δ, was set to 100 ms and the gradient pulse width, δ, was set to 8 ms. 32 spectra were collected for increasing gradient strengths from 2% to 95% of the maximum strength in a linear fashion. The resonance of Y25 in the unfolded state was used

for the analysis. The reported hydrodynamic radius of dioxane, 2.12 Å, was used to calculate the hydrodynamic radius of the protein [9].

2.11.6 ^{15}N R_2 relaxation experiments

^{15}N R_2 relaxation experiments were conducted on ^{15}N-labeled wild type NTL9 in 8.3 M urea on a Varian 600 MHz spectrometer equipped with a cryoprobe using a Carr-Purcell-Meiboom-Gill (CPMG) sequence. The 180˚ pulse spacing in the CPMG sequence was 1 ms. Spectra were collected in an interleaved manner with relaxation delays set to 14 (×2), 28, 38, 50(×2), 62, 86(×2), 98 and 110 ms. The spectral width was 8012.8 Hz (^{1}H) × 1920.0 Hz (^{15}N) with 1024 (^{1}H) × 512 (^{15}N) complex points. A recycle delay of 3 s was used. R_2 rates were determined using NMRViewJ to fit the peak intensities to two-parameter exponential decay. ^{15}N R_2 rates were analyzed using the phenomenological model of Schwalbe [60] by fitting the experimental R_2 rates to Equation 2.2.

$$R_2(i) = R_2(\text{int}) \sum_{j=1}^{N} \exp(-\frac{|i-j|}{\lambda}) \qquad (2.2)$$

$R_2(i)$ is the experimental R_2 value for residue i, $R_2(int)$ is the intrinsic relaxation rate which depends on the temperature and the viscosity of the solution. N is the total number of residues in the protein, and λ is the apparent persistence length of the chain.

2.11.7 Preparation of spin labeled samples

Samples were prepared by dissolving 3 mg of a NTL9 Cys mutant in 600 µL NMR buffer (20 mM sodium acetate, 100 mM sodium chloride). 12 µL of a 300 mM MTSL ((1-oxyl-2,2,5,5-tetramethyl-3-pyrroline-3-methyl)methanesulfonate) stock solution was added. The reaction was performed at room temperature for 10 h, and a Sephadex G25 column was used to remove excessive MTSL. For native state experiments, the protein sample was split into two equal

aliquots. For one of the aliquots, 300 μL NMR buffer was added (paramagnetic form); for the other aliquot, 270 μL of NMR buffer plus 30 μL of 100 mM TCEP (tris(2-carboxyethyl)phosphine) stock solution was added to reduce the protein (diamagnetic form). The pH was then adjusted to 5.5.

For the DSE studies, urea was added to a concentration of 10 M, (determined from the refractive index). The sample was split into two equal aliquots. For the diamagenetic form, 30 μL of 100 mM TCEP stock solution and NMR buffer was added; for the paramagnetic form, only NMR buffer was added. The final concentration of urea was 8.3 M for both samples, determined by measuring the refractive index.

Experiments for NMR assignments were conducted using a 500 MHz Bruker spectrometer with a cryoprobe. A Bruker 600 MHz spectrometer equipped with a cryoprobe was used for the Pulsed-field Gradient NMR Diffusion experiments. ^{1}H-^{15}N HSQC experiments were collected for both the paramagnetic form and diamagnetic form on a Varian Inova 600 MHz spectrometer with a conventional probe at 12 ˚C. The PRE $^{1}H_N$-Γ_2 rates measurement were performed in a two-time-point approach [46] on a Bruker 700 MHz spectrometer with a conventional probe.

2.11.8 Calculation of the theoretical PRE intensity ratios using the analytical Gaussian chain model and atomistic ensembles from the EV limit

A Gaussian random coil model and ensembles from EV limit simulations were used to calculate the theoretical intensity ratio and $^{1}H_N$-Γ_2 rates for a completely unstructured polypeptide. The Gaussian model assumes that the distance between each residue and the spin

label site follows a Gaussian distribution for the root-mean-square distance between residues [41, 61]:

$$< r^2 >= nl^2 \left(\frac{1+\alpha}{1-\alpha} - \frac{2\alpha(1-\alpha^n)}{n(1-\alpha)^2} \right) \tag{2.3}$$

where r is the distance between a residue and the spin label site, n is the number of residues between residue i and the spin label site, l is the link length of the chain, taken to be 3.8 Å, and α is the cosine of the bond-angle supplements for the freely rotating chain model, which was set to 0.8 based on experimentally determined estimates of statistical segment lengths in poly-L-alanine. The contribution of paramagnetic relaxation enhancement to the transverse relaxation rate, Γ_2 was calculated from Equation 2.4:

$$\Gamma_2 = \frac{K}{r^6}(4\tau_c + \frac{3\tau_c}{1+\omega_H^2\tau_c^2}) \tag{2.4}$$

Here, K is 1.23×10^{-32} cm^6s^{-2}, r is the distance between a given residue and the spin label site, ω_H is the proton Larmor frequency, and τ_C is the effective correlation time, which is 4 ns for NTL9 DSE in 8.3 M urea calculated from ^{15}N R_1 and R_2 relaxation rates. The peak intensity ratios between the paramagnetic and diamagnetic forms were calculated using Equation 2.5:

$$\frac{I_P}{I_D} = \frac{R_{2D}\exp(-\Gamma_2 t)}{R_{2D}+\Gamma_2} \tag{2.5}$$

Here, R_{2D} is the transverse relaxation rate of the backbone amide protons in the diamagnetic form of the DSE in 8.3 M urea. The average value was measured to be 14 s^{-1} using 1D NMR methods. The parameter t is the total duration of the INEPT delays, which is 12 ms for the HSQC pulse sequence.

2.11.9 Details of the Metropolis Monte Carlo (MC) simulations

We used the CAMPARI software package (http://campari.sourceforge.net/) employing

the ABSINTH implicit solvation model [47] and the underlying forcefield paradigm. Parameters were taken from the abs3.2_opls.prm set. NTL9 was modeled in atomic detail and the ABSINTH implicit solvation model was used to model solvent-mediated interactions.

The internal degrees of freedom included the backbone ϕ, ψ, ω and sidechain χ dihedral angles. Rigid-body moves simultaneously change rotational and translational degrees of freedom of the protein whereas translational moves were applied to alter the positions of mobile ions. The frequencies with which different moves were chosen are summarized in Table 2.1.

Move type	**Parameters**
Rigid-body	9% (50%, 10Å, 20º)
Random cluster	1% (50%, 10Å, 20º)
Concerted rotation	6.3%
Omega (ω)	5.67% (90%, 5º)
Sidechain (χ_1, …, χ_n)	27% (4x, 60%, 30º)
Backbone (ϕ, ψ)	51.03% (70%, 10º)

Table 2.1: Summary of MC move sets and frequencies used for all simulations

The first value listed in parentheses of row 2, column 2 of Table 2.1 is the fraction of moves assigned to finite perturbations whereas the remaining attempts fully randomize the corresponding degrees of freedom. The second and third values are the maximum displacements associated with translational and rotational moves for finite perturbations. Alterations to the ω angle involve random perturbations of randomly chosen angles. The two sets of values in parentheses of row 5, column 2 of Table 2.1 are the fraction of ω-moves that attempt a stepwise perturbation along with the maximum step-size. Sidechain moves perturb the χ-angles of a given sidechain in the peptide. In each attempt to alter sidechain degrees of the freedom, a random

number of χ-angles are given random orientations. Sidechain moves are inexpensive and therefore several sidechains are sampled simultaneous during each move. The number of sidechains sampled is the first value in parentheses of row 6, column 2 of Table 2.1. The remaining two values in parentheses give the fraction of χ-moves with a finite perturbation and the maximum value of that perturbation. Backbone moves simultaneously perturb the ϕ- and the ψ-angle of a given residue. The values in parentheses are interpreted the same way as for ω-moves. Concerted rotations simultaneously perturb eight consecutive backbone dihedral angles using the algorithm developed by Dinner [62]. This move set allows us to simultaneously probe multiple length scales simultaneously and efficiently.

2.11.10 Additional details regarding the generation of the starting conformation used in all simulations

The starting conformation for each of the 220 independent simulations was based on the coordinates deposited in the protein data bank, identifier 2HBB. In this model, the coordinates of five C-terminal residues were not resolved. These were constructed manually using the following procedure: The positions of all protein backbone atoms over the first 51 residues were constrained to their final position after the equilibration procedure described in the main text. All torsional degrees of freedom over the sidechains in the first 51 residues were also restrained as described in the main text. No constraints or restraints were applied over the last 5 residues and all backbone and sidechain degrees of freedom were sampled. The resulting system was subjected to 500,000 MC steps of sampling at T = 330 Kelvin followed by 1,000 steps of steepest descent minimization. The added segment adopts partially alpha-helical conformations and displays significant fraying of the C-terminal end. The heavy atom RMSD over the first 51

residues of 2HBB and the resulting structure was 1.14 Å. The final structure that includes all 56 residues was used as the starting conformation for all simulations.

2.11.11 The Monte Carlo sampling protocol

All results presented in this work were generated using the following thermal unfolding protocol: Ten independent MC simulations were used for each of the following temperatures T = [240, 260, 280, 290, 300, 310, 320, 330, 340, 345, 350, 355, 360, 365, 370, 375, 380, 390, 400, 430, 450, 500]. Additional sampling enhancements such as temperature replica exchange were not used because 1) the quality of sampling was sufficient without its use and 2) since each simulation was truly independent from all others it allowed evaluation of the reproducibility of all observables across independent simulations.

2.11.12 Calculation of the paramagnetic relaxation enhancement (PRE) intensity profile from simulation and comparison to experimental PRE values

The PRE intensity ratio ($I_{paramagnetic}$ / $I_{diamagnetic}$) profile was estimated from simulation for the following five sites in NTL9: K_2, K_{10}, K_{32}, A_{49}, K_{51}. Note that our simulations contain no mutations or spin labels on NTL9, which is in contrast to the experiments that include a nitroxide spin label. The pairwise distance distribution between the β-carbon *i* located at each of the five PRE sites and each backbone amide nitrogen atom *k* was accumulated every 500 MC steps with a bin size of 0.2Å. T-WHAM [(63)] was used to re-weight and combine these distributions collected at all 22 temperatures to maximally inform the distribution at each single target temperature. The average inverse sixth power of each *ik* pair at each target temperature is calculated as follows:

$$\left[\left\langle \frac{1}{R_{ik}^6} \right\rangle_{T=T_{target}}\right] = \frac{\sum_{t=1}^{n_{temps}} \sum_{j=1}^{n_{bins}} \left(\frac{1}{r_j^{ik}}\right)^6 \omega_j^{ik}(T_{target}, t) P(r_j^{ik}) \Delta r^{ik}}{\sum_{t=1}^{n_{temps}} \sum_{j=1}^{n_{bins}} \omega_j^{ik}(T_{target}, t)} \tag{2.6}$$

Here, $P(r_j^{ik})\Delta r^{ik}$ quantifies the probability that the distance r_j^{ik} assumes a value between $r_j^{ik} + \Delta r_j^{ik}$. In this case, $\Delta r_j^{ik} = 0.2$Å. $\omega_j^{jk}(T_{target}, t)$ is the T-WHAM derived weighting factor enabling the combination of distance distribution information across all simulated temperatures *t* to n_{temps} back to the ensemble at $T=T_{target}$. The contribution of paramagnetic relaxation enhancement to the transverse relaxation rate, R_{2P} was calculated using:

$$R_{2P} = \frac{K}{\left\langle R_{ik}^6 \right\rangle}\left(4\tau_c + \frac{3\tau_c}{1+\omega_H^2\tau_c^2}\right) \tag{2.7}$$

Here, *K* is 1.23×10^{-32} cm^6s^{-2}, *r* is the distance between a given residue and the spin label site. τ_C is the effective correlation time, which is 4 ns for NTL9 DSE in 8.3 M urea calculated from ^{15}N R_1 and R_2 relaxation rates and ω_H is the Larmor frequency of proton. Equation 2.8 reduces to:

$$R_{2P} = 1.9745\times10^8\left\langle \frac{1}{R_{ik}^6} \right\rangle \tag{2.8}$$

The peak intensity ratios between the paramagnetic and diamagnetic forms were calculated using:

$$\frac{I_p}{I_D} = \frac{R_{2D}\exp(-R_{2P}t)}{R_{2D}+R_{2P}} \tag{2.9}$$

Here, R_{2D} is the transverse relaxation rate of the backbone amide protons in the diamagnetic form of the DSE in 8.3 M urea. The average value was measured to be 14 s^{-1} using a 1D sequence and *t* is the total duration of the INEPT delays, which is 12 ms for the HSQC pulse sequence.

2.11.13 Alignment of bacterial N-terminal NTL9 sequences

Full-length bacterial NTL9 sequences were located in UniProt [(64)]. We sampled evenly for mesophiles, psychrophiles, and thermophiles. ClustalW [(65)] was used to align the full-length sequences and all settings were default: transition weighting = 0.5; alignment weight matrix = BLOSUM 62; gap opening penalty = 10; gap extension penalty = 0.1; end gap penalty = 0.5; gap distance = 1. Jalview [(66)] was used to display the N-terminal portion of this alignment.

2.12 References

1. Meng, W., Lyle, N., Pappu, R. V., and Raleigh, D. P. (2012) Unfolded Proteins Under Strongly Denaturing Conditions Form Long-Range Contacts In The Absence Of Significant Secondary Structure, *Proceedings of the National Academy of Sciences of the United States of America Submitted*.
2. Ziv, G., Thirumalai, D., and Haran, G. (2009) Collapse transition in proteins, *Physical Chemistry Chemical Physics 11*, 83-93.
3. Wu, Y., Kondrashkina, E., Kayatekin, C., Matthews, C. R., and Bilsel, O. (2008) Microsecond acquisition of heterogeneous structure in the folding of a TIM barrel protein, *Proceedings of the National Academy of Sciences of the United States of America 105*, 13367-13372.
4. Dobson, C. M. (2003) Protein folding and misfolding, *Nature 426*, 884-890.

5. Jahn, T. R., and Radford, S. E. (2008) Folding versus aggregation: Polypeptide conformations on competing pathways, *Archives of Biochemistry and Biophysics 469*, 100-117.

6. Zhou, H. X., Rivas, G. N., and Minton, A. P. (2008) Macromolecular crowding and confinement: Biochemical, biophysical, and potential physiological consequences, *Annual Review of Biophysics 37*, 375-397.

7. Cho, J. H., and Raleigh, D. P. (2006) Electrostatic interactions in the denatured state and in the transition state for protein folding: Effects of denatured state interactions on the analysis of transition state structure, *Journal of Molecular Biology 359*, 1437-1446.

8. Tanford, C. (1968) Protein Denaturation, *Adv. Protein Chem. 23*, 121-282.

9. Wilkins, D. K., Grimshaw, S. B., Receveur, V., Dobson, C. M., Jones, J. A., and Smith, L. J. (1999) Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques, *Biochemistry 38*, 16424-16431.

10. Kohn, J. E., Millett, I. S., Jacob, J., Zagrovic, B., Dillon, T. M., Cingel, N., Dothager, R. S., Seifert, S., Thiyagarajan, P., Sosnick, T. R., Hasan, M. Z., Pande, V. S., Ruczinski, I., Doniach, S., and Plaxco, K. W. (2004) Random-coil behavior and the dimensions of chemically unfolded proteins, *Proceedings of the National Academy of Sciences of the United States of America 101*, 12491-12496.

11. Schäfer, L. (1999) *Excluded Volume Effects in Polymer Solutions as Explained by the Renormalization Group*, Springer, Berlin.

12. Hofmann, H., Soranno, A., Borgia, A., Gast, K., Nettels, D., and Schuler, B. (2012) Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with

single-molecule spectroscopy, *Proceedings of the National Academy of Sciences of the United States of America 109*, 16155-16160.

13. Jha, A. K., Colubri, A., Freed, K. F., and Sosnick, T. R. (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*, 13099-13104.

14. Bernado, P., Blanchard, L., Timmins, P., Marion, D., Ruigrok, R. W. H., and Blackledge, M. (2005) A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering, *Proceedings of the National Academy of Sciences of the United States of America 102*, 17002-17007.

15. Meier, S., Blackledge, M., and Grzesiek, S. (2008) Conformational distributions of unfolded polypeptides from novel NMR techniques, *J. Chem. Phys. 128*, 052204.

16. Salmon, L., Nodet, G., Ozenne, V., Yin, G. W., Jensen, M. R., Zweckstetter, M., and Blackledge, M. (2010) NMR Characterization of Long-Range Order in Intrinsically Disordered Proteins, *Journal of the American Chemical Society 132*, 8407-8418.

17. Tran, H. T., and Pappu, R. V. (2006) Toward an accurate theoretical framework for describing ensembles for proteins under strongly denaturing conditions, *Biophysical Journal 91*, 1868-1886.

18. Tran, H. T., Wang, X. L., and Pappu, R. V. (2005) Reconciling observations of sequence-specific conformational propensities with the generic polymeric behavior of denatured proteins, *Biochemistry 44*, 11369-11380.

19. Mao, A. H., Crick, S. L., Vitalis, A., Chicoine, C. L., and Pappu, R. V. (2010) Net charge per residue modulates conformational ensembles of intrinsically disordered proteins,

Proceedings of the National Academy of Sciences of the United States of America 107, 8183-8188.

20. Dedmon, M. M., Lindorff-Larsen, K., Christodoulou, J., Vendruscolo, M., and Dobson, C. M. (2005) Mapping long-range interactions in alpha-synuclein using spin-label NMR and ensemble molecular dynamics simulations, *Journal of the American Chemical Society 127*, 476-477.

21. Ding, F., Jha, R. K., and Dokholyan, N. V. (2005) Scaling behavior and structure of denatured proteins, *Structure 13*, 1047-1054.

22. Zhou, H. X. (2004) Polymer models of protein stability, folding, and interactions, *Biochemistry 43*, 2141-2154.

23. Rubinstein, M., and R.H., C. (2003) *Polymer Physics*, Oxford University Press, New York, USA.

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

25. Francis, C. J., Lindorff-Larsen, K., Best, R. B., and Vendruscolo, M. (2006) Characterization of the residual structure in the unfolded state of the Delta 131 Delta fragment of staphylococcal nuclease, *Proteins-Structure Function and Bioinformatics 65*, 145-152.

26. Kristjansdottir, S., Lindorff-Larsen, K., Fieber, W., Dobson, C. M., Vendruscolo, M., and Poulsen, F. M. (2005) Formation of native and non-native interactions in ensembles of denatured ACBP molecules from paramagnetic relaxation enhancement studies, *J. Mol. Biol. 347*, 1053-1062.

27. Lindorff-Larsen, K., Kristjansdottir, S., Teilum, K., Fieber, W., Dobson, C. M., Poulsen, F. M., and Vendruscolo, M. (2004) Determination of an ensemble of structures representing the denatured state of the bovine acyl-coenzyme a binding protein, *J. Am. Chem. Soc. 126*, 3291-3299.

28. Robustelli, P., Kohlhoff, K., Cavalli, A., and Vendruscolo, M. (2010) Using NMR Chemical Shifts as Structural Restraints in Molecular Dynamics Simulations of Proteins, *Structure 18*, 923-933.

29. Fieber, W., Kristjansdottir, S., and Poulsen, F. M. (2004) Short-range, long-range and transition state interactions in the denatured state of ACBP from residual dipolar couplings, *Journal of Molecular Biology 339*, 1191-1199.

30. Mayor, U., Grossmann, J. G., Foster, N. W., Freund, S. M. V., and Fersht, A. R. (2003) The denatured state of engrailed homeodomain under denaturing and native conditions, *Journal of Molecular Biology 333*, 977-991.

31. Felitsky, D. J., Lietzow, M. A., Dyson, H. J., and Wright, P. E. (2008) Modeling transient collapsed states of an unfolded protein to provide insights into early folding events, *Proceedings of the National Academy of Sciences of the United States of America 105*, 6278-6283.

32. Mohana-Borges, R., Goto, N. K., Kroon, G. J. A., Dyson, H. J., and Wright, P. E. (2004) Structural characterization of unfolded states of apomyoglobin using residual dipolar couplings, *Journal of Molecular Biology 340*, 1131-1142.

33. Marsh, J. A., Neale, C., Jack, F. E., Choy, W. Y., Lee, A. Y., Crowhurst, K. A., and Forman-Kay, J. D. (2007) Improved structural characterizations of the drkN SH3 domain

unfolded state suggest a compact ensemble with native-like and non-native structure, *Journal of Molecular Biology 367*, 1494-1510.

34. Klein-Seetharaman, J., Oikawa, M., Grimshaw, S. B., Wirmer, J., Duchardt, E., Ueda, T., Imoto, T., Smith, L. J., Dobson, C. M., and Schwalbe, H. (2002) Long-range interactions within a nonnative protein, *Science 295*, 1719-1722.

35. Tran, H. T., Mao, A., and Pappu, R. V. (2008) Role of backbone - Solvent interactions in determining conformational equilibria of intrinsically disordered proteins, *J. Am. Chem. Soc. 130*, 7380-7392.

36. Voelz, V. A., Bowman, G. R., Beauchamp, K., and Pande, V. S. (2010) Molecular Simulation of ab Initio Protein Folding for a Millisecond Folder NTL9(1-39), *Journal of the American Chemical Society 132*, 1526-1528.

37. Sherman, E., and Haran, G. (2006) Coil-globule transition in the denatured state of a small protein, *Proceedings of the National Academy of Sciences of the United States of America 103*, 11539-11543.

38. Mueller-Spaeth, S., Soranno, A., Hirschfeld, V., Hofmann, H., Rueegger, S., Reymond, L., Nettels, D., and Schuler, B. (2010) Charge interactions can dominate the dimensions of intrinsically disordered proteins, *Proceedings of the National Academy of Sciences of the United States of America 107*, 14609-14614.

39. Schwarzinger, S., Kroon, G. J. A., Foss, T. R., Wright, P. E., and Dyson, H. J. (2000) Random coil chemical shifts in acidic 8 M urea: Implementation of random coil shift data in NMRView, *Journal Of Biomolecular Nmr 18*, 43-48.

40. Marsh, J. A., Singh, V. K., Jia, Z. C., and Forman-Kay, J. D. (2006) Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: Implications for fibrillation, *Protein Science 15*, 2795-2804.

41. Lietzow, M. A., Jamin, M., Dyson, H. J., and Wright, P. E. (2002) Mapping long-range contacts in a highly unfolded protein, *Journal of Molecular Biology 322*, 655-662.

42. Neri, D., Billeter, M., Wider, G., and Wuthrich, K. (1992) NMR determination of residual structure in a urea-denatured protein, the 434-repressor, *Science 257*, 1559-1563.

43. Schmidt, P. G., and Kuntz, I. D. (1984) Distance measurements in spin-labeled lysozyme, *Biochemistry 23*, 4261-4266.

44. Gillespie, J. R., and Shortle, D. (1997) Characterization of long-range structure in the denatured state of staphylococcal nuclease .1. Paramagnetic relaxation enhancement by nitroxide spin labels, *Journal of Molecular Biology 268*, 158-169.

45. Xue, Y., Podkorytov, I. S., Rao, D. K., Benjamin, N., Sun, H. L., and Skrynnikov, N. R. (2009) Paramagnetic relaxation enhancements in unfolded proteins: Theory and application to drkN SH3 domain, *Protein Science 18*, 1401-1424.

46. Iwahara, J., Tang, C., and Clore, G. M. (2007) Practical aspects of (1)H transverse paramagnetic relaxation enhancement measurements on macromolecules, *Journal of Magnetic Resonance 184*, 185-195.

47. Vitalis, A., and Pappu, R. V. (2009) ABSINTH: A New Continuum Solvation Model for Simulations of Polypeptides in Aqueous Solutions, *J. Comput. Chem. 30*, 673-699.

48. Kaminski, G., Friesner, R., Tirado-Rives, J., and Jorgensen, W. (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides, *Journal of Physical Chemistry B 105*, 6487.

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

50. Breton, M. D., Devore, M. D., and Brown, D. E. (2008) A tool for systematically comparing the power of tests for normality, *J. Stat. Comput. Simul. 78*, 623-638.

51. Ha, B. Y., and Thirumalai, D. (1992) Conformations of a polyelectrolyte chain, *Physical Review A 46*, R3012-R3015.

52. Cellmer, T., Henry, E. R., Hofrichter, J., and Eaton, W. A. (2008) Measuring internal friction of an ultrafast-folding protein, *Proceedings of the National Academy of Sciences of the United States of America 105*, 18320-18325.

53. Clementi, C., and Plotkin, S. S. (2004) The effects of nonnative interactions on protein folding rates: Theory and simulation, *Protein Science 13*, 1750-1766.

54. Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J., and Svergun, D. I. (2003) PRIMUS: a Windows PC-based system for small-angle scattering data analysis, *J Appl Crystallogr 36*, 1277-1282.

55. Guinier, A. (1955) *Small-Angle X-ray Scattering*, John Wiley, New York.

56. Delaglio, F., Grzesiek, S., Vuister, G., Zhu, G., Pfeifer, J., and Bax, A. (1995) NMRPipe: A multidimensional spectral processing system based on UNIX pipes, *Journal of Biomolecular NMR 6*, 277-293.

57. Johnson, B. A. (2004) Using NMRView to visualize and analyze the NMR spectra of macromolecules, *Methods in molecular biology (Clifton, N.J.) 278*, 313-352.

58. Wishart, D. S., Bigam, C. G., Holm, A., Hodges, R. S., and Sykes, B. D. (1995) 1H, 13C and 15N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects, *J Biomol NMR 5*, 67-81.

59. Schwarzinger, S., Kroon, G. J., Foss, T. R., Chung, J., Wright, P. E., and Dyson, H. J. (2001) Sequence-dependent correction of random coil NMR chemical shifts, *J Am Chem Soc 123*, 2970-2978.

60. Schwalbe, H., Fiebig, K. M., Buck, M., Jones, J. A., Grimshaw, S. B., Spencer, A., Glaser, S. J., Smith, L. J., and Dobson, C. M. (1997) Structural and dynamical properties of a denatured protein. Heteronuclear 3D NMR experiments and theoretical simulations of lysozyme in 8 M urea, *Biochemistry 36*, 8977-8991.

61. Sung, Y. H., and Eliezer, D. (2007) Residual structure, backbone dynamics, and interactions within the synuclein family, *Journal of Molecular Biology 372*, 689-707.

62. Dinner, A. R. (2000) Local deformations of polymers with nonplanar rigid main-chain internal coordinates, *Journal of Computational Chemistry 21*, 1132-1144.

63. Chodera, J. D., Swope, W. C., Pitera, J. W., Seok, C., and Dill, K. A. (2007) Use of the Weighted Histogram Analysis Method for the Analysis of Simulated and Parallel Tempering Simulations, *Journal of Chemical Theory and Computation 3*, 26-41.

64. Consortium, T. U. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt), *Nucleic Acids Research 40*, D71-D75.

65. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007) Clustal W and Clustal X version 2.0, *Bioinformatics 23*, 2947-2948.

66. Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., and Barton, G. J. (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench, *Bioinformatics 25*, 1189-1191.

Chapter 3

Thermodynamics of β-sheet formation in polyglutamine

This chapter is adapted from an article [1] in Biophysical Journal co-authored by Andreas Vitalis (AV) and Nicholas J. Lyle (NJL). AV established the protocol for biased simulations. NL tested and optimized the protocol and performed the simulations. NL and AV interpreted the results. This work was supported by grant 5R01-NS056114 from the National Institutes of Health.

3.1 Introduction

Nine neurodegenerative diseases are associated with polyglutamine expansion mutations [2]. Ages of onset and severity of disease at onset are inversely correlated with the lengths of polyglutamine expansions [3]. Proteolysis of mutant proteins releases fragments rich in polyglutamine that aggregate to form insoluble neuronal intranuclear inclusions [4]. Aggregates are fibrillar and amyloid-like [5] with high β-sheet contents. Based on fiber diffraction data, Perutz proposed a model in which the individual polyglutamine molecules within aggregates are arranged in a β-helical conformation [6]. The fiber diffraction data have also been shown to be consistent with flat β-sheet architectures for individual polyglutamine peptides [7].

In contrast to structures adopted by individual polyglutamine molecules in fibrillar aggregates, monomeric polyglutamine molecules have spectroscopic signals that indicate a lack of well-defined secondary structures [8, 9]. It has also been shown that monomeric polyglutamine forms collapsed structures in water [10], and this observation is consistent with water being a poor solvent for polyglutamine [11]. Monomeric polyglutamine is intrinsically disordered, irrespective

of chain length, and these molecules sample heterogeneous ensembles of collapsed structures in water.

Chen *et al*. [12] proposed a model to connect intrinsically disordered monomers and β-sheet structures that are prominent in fibrillar forms. Analysis of kinetic data using a homogeneous nucleation model yielded a nucleus size of one [12]. The results were interpreted as follows: Monomeric polyglutamine is in a pre-equilibrium between a folded, toxic, β-sheet structure and the disordered ensemble. Whenever the toxic fold is adopted, the thermodynamically unfavorable nucleus is populated and the resulting ordered conformation is elongated through monomer addition. A schematic of this proposal is shown in Figure 3.1. The proposal put forth by Chen *et al*. is difficult to test experimentally because the critical nucleus is, by definition, a rare species.

Recently [13], we corroborated experimental results that polyglutamine chains spanning the pathological threshold range for Huntington's disease ($N \approx 37$) adopt disordered, compact conformations at the monomer level. We also showed that they associate spontaneously to form disordered dimers. These studies were carried out using the ABSINTH implicit solvation model combined with the all atom OPLS-AA/L forcefield [13, 14]. Here we use the same combination of forcefield and implicit solvation model to answer a set of questions and test all aspects of the homogeneous nucleation model proposed by the analysis of Wetzel and coworkers [12, 15]. The questions of interest are as follows: What is the probability that monomeric polyglutamine forms structures with high β-content under ambient conditions? Are such species metastable states along a reaction coordinate that measures the net β-content of the chain? How does the likelihood of forming species high in β-content vary with chain length? If we bias individual polyglutamine chains toward structures with high β-content, do we observe a change in the

spontaneity of dimer formation? Finally, do our results support the idea of a structure-driven, homogeneous nucleation pathway for polyglutamine aggregation? The rest of this manuscript is organized as follows: First we introduce the details of our methodology. This is followed by a presentation of results that allow us to answer the questions raised above and we conclude with a summary and discussion of our results.

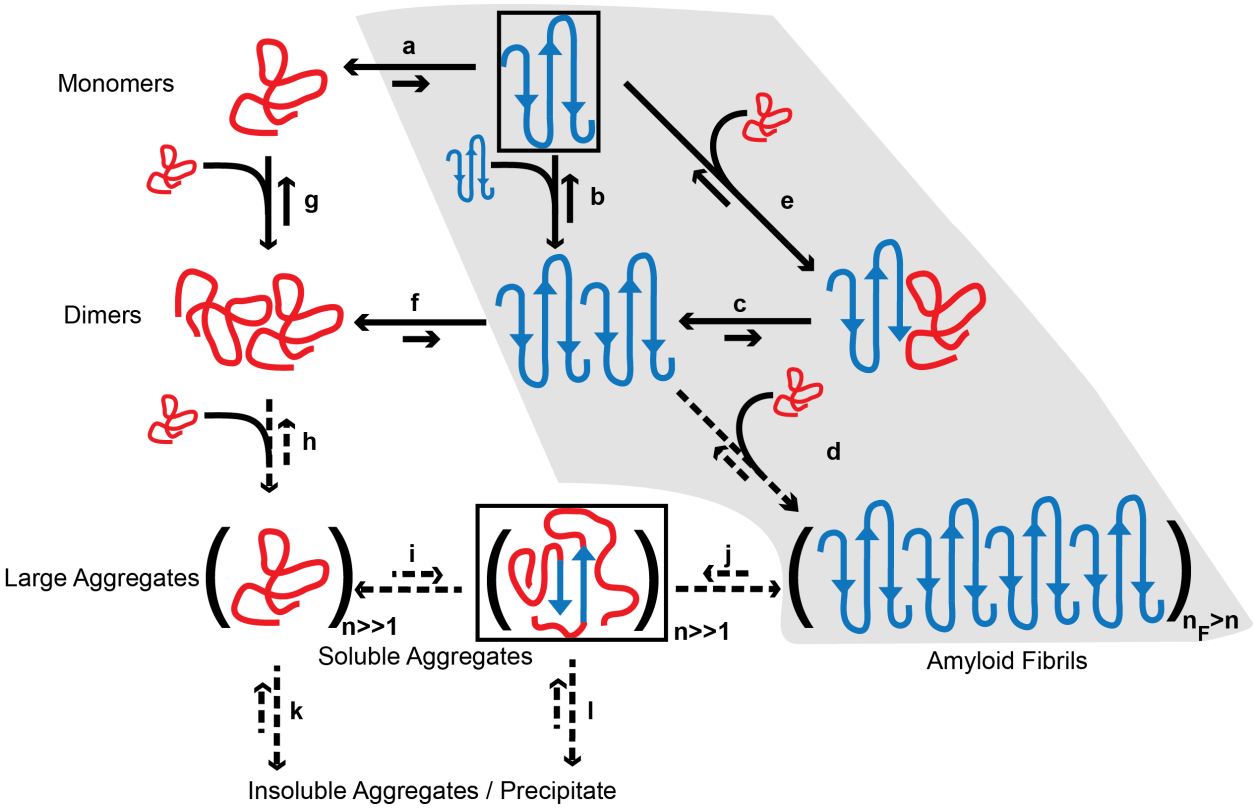


Figure 3.1: Schematic of possible aggregation pathways for polyglutamine *in vitro*. n denotes the number of polyglutamine molecules within a disordered aggregate and n_F denotes the number of polyglutamine molecules within an ordered amyloid fibril. The ordered amyloid fibril rich in β-sheets is shown in the bottom right corner of the schematic. The gray shaded region encompasses steps (a), (e), (c), and (d) and depicts the homogeneous nucleation proposal of Chen et al. [12]. We investigated the thermodynamics of step (a), which indicates that the formation of ordered conformations is thermodynamically unfavorable. Monomeric

polyglutamine prefers disordered, collapsed conformations, left of step (a). Step (b) pertains to the thermodynamics of interactions between chains that have been restrained to adopt ordered conformations. Associativities of restrained chains – step (b) – are akin to the associativities of unrestrained chains – step (g). However, the likelihood that chains will sample the associations shown in step (b) is very small because this is tied to the equilibria in step (a), which requires the population of the conformations with high β-content, and Figure 3.3 shows that this is highly unlikely. Similarly, step (f) shows that disordered dimers are thermodynamically favored to dimers with high β-contents in individual chains. This is the result of linkage to step (a) as discussed above. The aggregates achieved in step (h) are likely to be large (in terms of n) and exhibit spherical, "liquid-like" [16-18] organization of polyglutamine chains around each other. Step (i) depicts a slow conformational conversion of individual / small numbers of chains to β-sheets. This slow step is likely to lead to the creation of an ordered template for fibril formation via monomer or oligomer addition and elongation to yield the ordered amyloid fibril. Steps (a), (b), and (g) are anchored in the collection of data generated in this work and previous studies. However, the reversible associations depicted in step (h) and the conformational conversions depicted in step (i) are yet to be tested.

3.2 Materials and Methods

3.2.1 Degrees of freedom and molecular forcefield

We simulated one or two polyglutamine molecules N-acetyl-$(Gln)_N$-N′-Methylamide; the chains were modeled in atomic detail and the number of glutamine residues was set to N = 5, 15, 30, and 45, respectively. Chains with N glutamine residues are denoted as Q_N. Markov chain Metropolis Monte Carlo (MC) Simulations were performed in the canonical ensemble and molecules were enclosed in a spherical droplet of radius 200Å, which was enforced using a stiff

harmonic boundary potential. The degrees of freedom were the backbone ϕ, ψ, ω and sidechain χ dihedral angles. For MC simulations with two chains, rigid-body coordinates, namely center-of-mass translations and rotations were included as additional degrees of freedom. Bond lengths and bond angles were held fixed at values prescribed by Engh and Huber [19].

We used parameters from the OPLS-AA/L forcefield [20] with appropriate modifications and the ABSINTH implicit solvent model [14]. The energy function is:

$$E_{\text{total}} = W_{\text{solv}} + W_{\text{el}} + U_{\text{LJ}} + U_{\text{tor}} \tag{3.1}$$

Here, W_{solv} is the direct mean field interaction (DMFI) term that captures the transfer of a polypeptide solute in a specific conformation from the gas phase into the continuum solvent. W_{el} denotes the mean field electrostatic term; the dielectric constant of the solvent is set to be ε=78. U_{LJ} models van der Waals interactions using the Lennard-Jones (LJ) model. Parameters for the LJ hard sphere radii and well depths are based on heats of fusion data for model compounds and are different from the choices made in standard forcefields. This choice allows us to omit many of the torsional potentials because the rotational barriers and minima are captured by excluded volume interactions alone [14]. U_{tor} represents torsional terms applied to dihedral angles that are cannot be captured by U_{LJ}. For polyglutamine, we use torsional potentials taken from OPLS-AA/L to maintain peptide dihedral angles (ω) in predominantly *trans*-configurations.

Salient features of the interplay between the W_{solv} and W_{el} terms are summarized as follows: Polypeptide chains are decomposed into a set of distinct solvation groups; for capped polyglutamine the solvation groups are N+1 backbone secondary amides and N sidechain primary amides. W_{solv} is a sum of contributions from each solvation group and for each of these groups we use experimentally measured free energies of solvation of appropriate model compound analogs as references for fully solvated states. The degree of solvent accessibility

modulates the DMFI and consequently W_{solv} varies with conformation. Solvent accessible volume fractions are used as the metric for solvent accessibility and this is used to evaluate the solvation states υ_k^i for atom *k* in solvation group *i*. Electrostatic interactions between non-bonded solute atoms with partial charges are fully screened by the continuum dielectric if the atoms are fully exposed to solvent. The screening of polar interactions is conformation dependent and the solvation states of atoms υ_k^i determine the extent to which the screening of electrostatic interactions is modulated. ABSINTH is built upon the distinct strengths of the generalized Born (GB) [21-23] and EEF1 [24] models. In accord with EEF1, the process of transferring solutes from the gas phase into the continuum solvent is treated in "one shot" without attempting to parse the distinct contributions from the polar and non-polar components. However, ABSINTH deviates from EEF1 in the way electrostatic interactions between solute atoms are handled. No explicit or implicit distance dependence is assumed for the dielectric response. Instead, solvation states of individual atoms υ_k^i are used to determine the extent to which the screening of electrostatic interactions is to be modulated by the protein environment. ABSINTH therefore captures the main strengths of the GB model while retaining the efficiency of the EEF1 model and it may be viewed as an effective interpolation between GB and EEF1.

3.2.2 Sampling methodology

Move sets for MC simulations included pivots about random backbone torsions, perturbations about random dihedrals, randomized sidechain rotations, and randomized changes to rigid body coordinates. Details regarding the move sets are summarized in Table 3.1. One of our goals was to quantify free energy penalties associated with sampling of β-rich conformations for monomeric polyglutamine. We defined a reaction coordinate f_β to assess global β-content within a molecule and used umbrella sampling to enhance the sampling of the low-likelihood

regions of conformational space. Results of independent umbrella sampling calculations were stitched together using weighted histogram analysis methods (WHAM) [25, 26] to generate unbiased potentials of mean force (PMFs) as a function of f_β, which is defined as

$$f_\beta = \frac{1}{N}\sum_{i=1}^{N} f_\beta^{(i)} \text{ where:}$$

$$f_\beta^{(i)} = \begin{cases} 1.0 & \text{if } (\phi_i, \psi_i) \text{ belongs to the } \beta\text{-basin} \\ \exp\left(-\tau_\beta d_{(i)}^2\right) & \text{otherwise} \end{cases} \tag{3.2}$$

$$d_{(i)}^2 = \left\{ \left(\sqrt{\left[\left(\phi_i - \phi_\beta\right) \bmod 2\pi\right]^2 + \left[\left(\psi_i - \psi_\beta\right) \bmod 2\pi\right]^2} - r_\beta \right) \bmod 2\pi \right\}^2 \tag{3.3}$$

Here, N is the number of glutamine residues in the sequence, mod2π terms correct for periodicity effects associated with distance calculations in angular space; (ϕ_β,ψ_β) define the reference (ϕ,ψ) values for an individual residue. If residue i adopts ϕ and ψ angles that lie within a circle of radius r_β, then the parameter $f_\beta^{(i)}$ is set to unity; otherwise, $f_\beta^{(i)}$ has a value between 0 and 1; the precise value is determined by two parameters, namely $d_{(i)}^2$ and τ_β. The latter is the width of the Gaussian function and ensures a continuous function. We used: $(\phi_\beta,\psi_\beta) = (-152°, 142°)$, r_β=50°, and τ_β=0.002 deg^{-2}.

Table 3.1: Overview of the frequency of the different Monte Carlo moves sets used in simulations of monomeric and pairs of polyglutamine molecules. To be able to probe multiple length scales simultaneously, Monte Carlo moves in ABSINTH either fully randomize a given degree of freedom, or perform a stepwise perturbation that has a maximum size. This is explained in detail in the footnotes. The frequencies for different moves are chosen to reflect the relevance of the various degrees of freedom to both the conformational equilibria and the

association of these peptides. Additionally, these choices reflect the associated computational cost. As an example, ω-angles are sampled relatively infrequently, as their values are expected to remain close to the perfect *trans*-conformation. Note that a small number of moves for each simulation were swap attempts for replica exchange.

Move Type	**Frequency of move sets for simulations of monomeric polyglutamine**	**Frequency of move sets for simulations with pairs of polyglutamine molecules**
Rigid-body[1]	-	30% (50%, 10Å, 20°)
Omega[2]	7% (90%, 5°)	4.9% (90%, 5°)
Sidechain[3]	30% (4x, 60%, 30°)	21% (4x, 40%, 30°)
Pivot[4]	63% (70%, 10°)	44.1% (70%, 10°)

1. Rigid-body moves simultaneously change rotational and translational degrees of freedom of the whole molecule. The first value listed in parentheses is the fraction of moves assigned to finite perturbations, whereas the remaining attempts fully randomize the respective degrees of freedom. The second and third values are the maximum translational and rotational step-sizes associated with the finite perturbations.
2. Moves that perturb the ω-angles of peptide units. Due to the acetyl and methylamide capping groups there are N+1 ω-angles for a chain length of N. The two sets of values in parentheses are the fraction of ω-moves, which attempt a stepwise perturbation along with the maximum step-size.
3. Sidechain moves perturb the χ-angles of a given sidechain in the peptide. In each attempt to alter sidechain degrees of the freedom, two of the three χ-angles are randomly altered. Sidechain moves are inexpensive and therefore several sidechains are sampled during each "move" (first value in parentheses). The remaining two values in parentheses again

give the fraction of χ-moves with a finite perturbation and the maximum value of that perturbation.

We assessed the validity of f_β as a measure of β-content by quantifying its ability to estimate β-content in proteins of known three-dimensional structures. We used PDBSelect [27] to create a database of 3,693 non-redundant protein structures from the protein data bank. Sequences in this dataset have less than 25% sequence identity with each other. For each structure in the dataset, we calculated the f_β values and their DSSP-E score [28], normalized by the number of residues, as an alternative to measure the degree of ordered β-sheet. Figure 3.2 shows the correlation between f_β and fractional DSSP-E scores. The correlation coefficient is 0.83. Figure 3.3 shows ribbon drawings of three-dimensional structures for five structures from the database. Since we do not have definitive prior knowledge of the type of ordered β-sheets that polyglutamine molecules adopt in fibrillar aggregates, we used f_β instead of fractional DSSP-E scores as a reaction coordinate for assessing the bias toward structures with high or low β-contents.

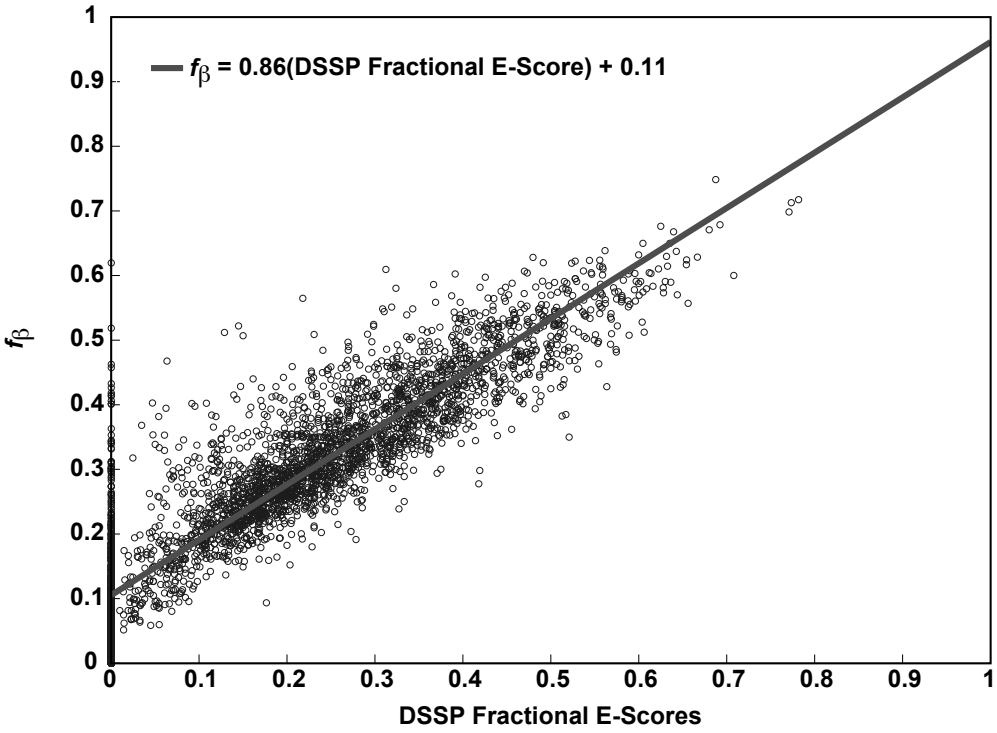


Figure 3.2: Correlation between fractional DSSP-E scores and f_β. The solid line is the line of best fit that quantifies the strength and direction of the linear correlation between f_β values and fractional DSSP-E scores. Parameters for the slope and intercept are shown in the inset. Structures that have high fractional DSSP-E scores also have high f_β values, although there is some scatter about the line of best fit. For approximately 27% of the structures in the dataset, the fractional DSSP-E scores are zero. Although the f_β values for most of these structures are small (≤ 0.3), they span a finite range of f_β values.

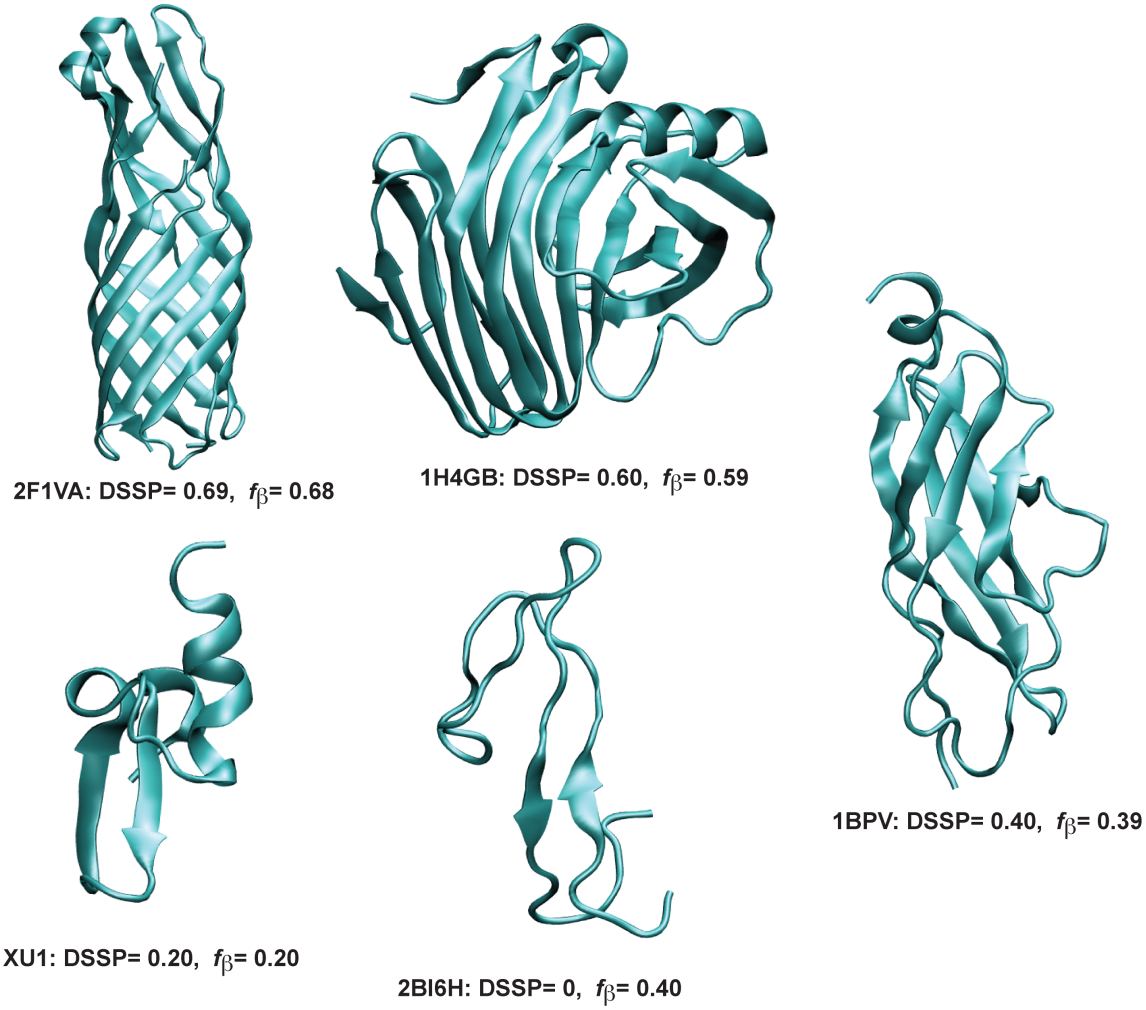


Figure 3.3: Representative structures from the PDBSelect database showing the correlation between f_β values and fractional DSSP-E scores. For each structure, the label denotes the PDB code with chain ID, fractional DSSP-E score, and f_β value, respectively. The correlation becomes weak when characteristic hydrogen-bonded patterns are absent within a structure as shown for 2BI6H.

Simulations for monomeric polyglutamine were carried out at 298K. For each chain length, we performed seventeen sets of distinct umbrella sampling simulations and in each simulation, f_β was restrained to one of seventeen target f_β^0 values: [0.0, 0.1, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.7, 0.75, 0.8, 0.9, 0.95, 1.0]. Each simulation was biased toward a

target value f_β^0 using a harmonic restraint potential $U_{restr} = k(f_\beta - f_\beta^0)^2$ that was added to the energy function shown in Equation 3.1. The number of (ϕ,ψ) pairs that contribute to f_β increases with N and the value of k varied with N. We used values of k=25 kcal/mol and 75 kcal/mol for Q_5 and Q_{15}, and k=150 kcal/mol for Q_{30} and Q_{45}, respectively. Therefore, k varies from 1.7 kcal/mol (Q_{45}) to 2.5 kcal/mol (Q_5, Q_{15}, and Q_{30}) per restrained degree of freedom.

For each window, the starting conformation was extracted at random from an ensemble of self-avoiding random walks. For Q_5, Q_{15}, and Q_{30} the first 10^6 MC steps were used for equilibration followed by 4×10^7 steps of production. For Q_{45}, we used 1.5×10^6 steps of equilibration and 6×10^7 steps of production. Sampling was enhanced using replica exchange MC in f_β-space [29]. For each chain length, we performed three independent replica exchange umbrella sampling MC runs. The quality of sampling was assessed by computing statistics for the extent of overlap of f_β histograms between adjacent windows and statistics for replica exchange. These details are documented in Figures 4 – 6.

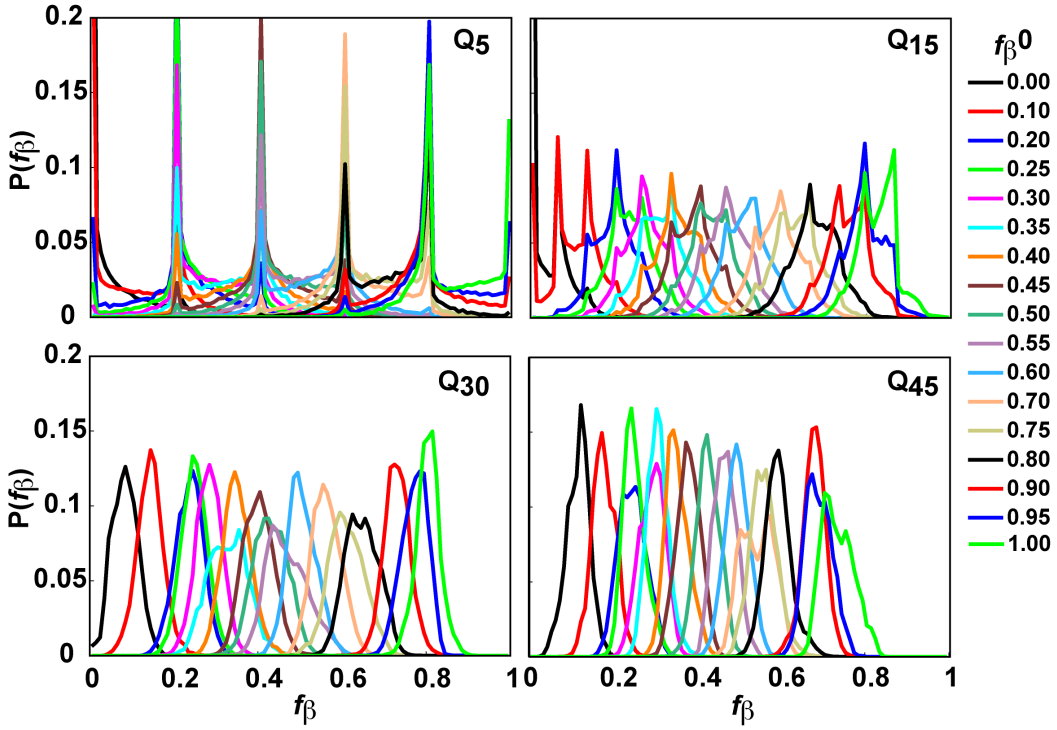


Figure 3.4: Distributions of f_β values obtained from the umbrella sampling simulations. The legend on the right identifies the target f_β^0 value for each simulation. Each histogram is the average from three independent simulations. The histograms indicate that the overlap between adjacent windows is finite and significant, thereby validating the f_β schedule and choice of restraints used in the umbrella sampling protocol. To further demonstrate the validity of our protocol, Figure 3.5 shows a summary of the overlap statistics extracted from the plots shown in this figure. As a reminder, we present data from umbrella sampling simulations that were performed using the following schedule of seventeen values for f_β^0: [0.0, 0.1, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.7, 0.75, 0.8, 0.9, 0.95 1.0].

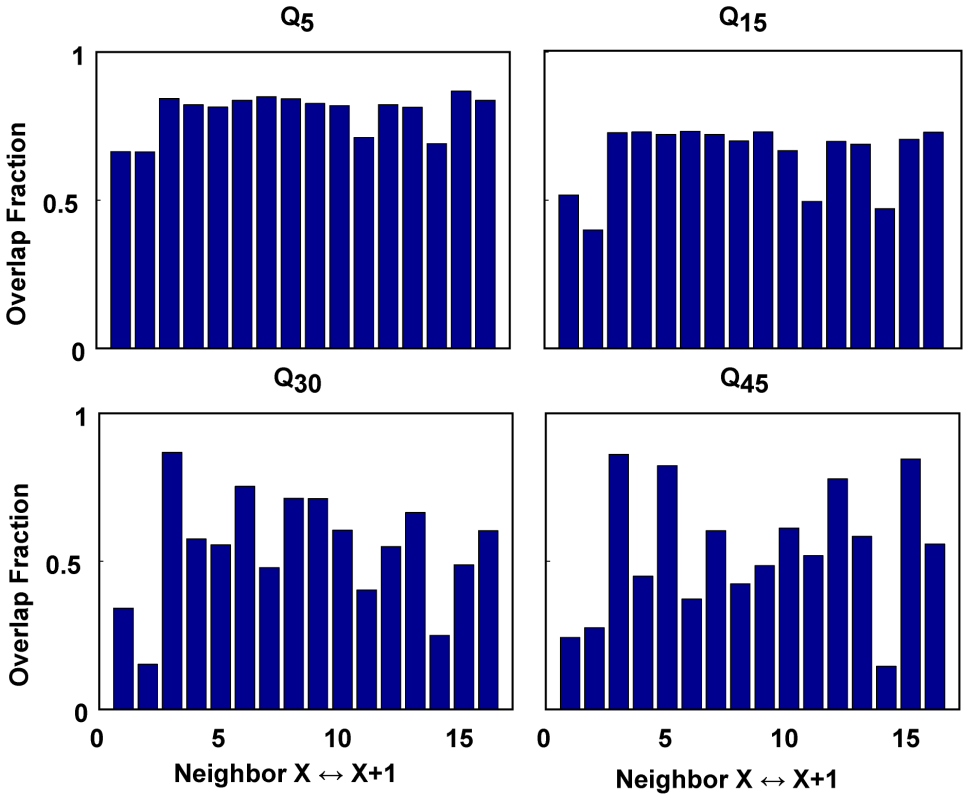


Figure 3.5: Quantification of overlap between adjacent f_β histograms shown in Figure 3.4.

Sixteen bars are shown for overlap statistics computed between seventeen sets of adjacent windows. The f_β schedule is as shown in the legend to Figure 3.4. For a pair of windows X and X+1, $\text{Overlap Fraction} = \dfrac{2 - \int_{f_\beta=0}^{f_\beta=1} \left| P_X\left(f_\beta\right) - P_{X+1}\left(f_\beta\right) \right| df_\beta}{2}$. We computed each integral numerically. In this formula, P_X and P_{X+1} are the average f_β histograms for windows X and X+1, respectively and these histograms are shown in Figure 3.4. If adjacent histograms P_X and P_{X+1} overlap perfectly, then the value of the overlap fraction is unity. Conversely, if the histograms do not overlap at all, then the overlap fraction is zero. The smallest overlap fraction values are seen for two pairs of neighbors in the simulations for Q_{30} and Q_{45}. However, the reconstructed PMFs obtained using either WHAM or TI-WHAM are smooth even in regions where the overlap is the smallest, although the error bars (standard errors and bootstrap errors) are large in these high free

energy regions. Overall, the quality of data obtained using the umbrella sampling protocol adopted in this work appears to be satisfactory and yields reliable PMFs thereby allowing us to draw the conclusions summarized in the main text.

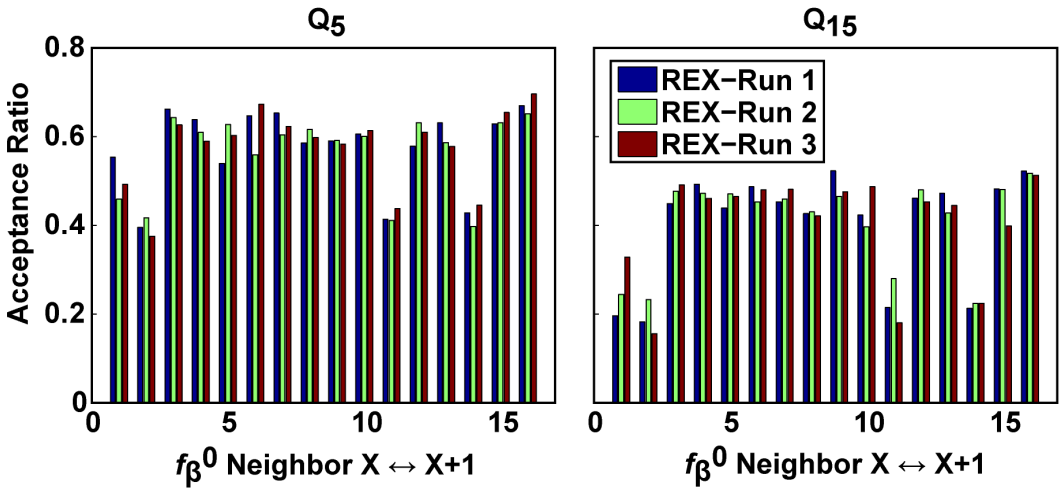


Figure 3.6: Plot of acceptance ratios for replica exchange swaps between nearest-neighbor windows in the umbrella sampling simulations that were carried out for monomeric polyglutamine. As noted in the text, we carried out three independent replica exchange simulations, which were combined with the umbrella sampling protocol. The f_β schedule used in these simulations is identical to that shown in the legend to Figure 3.4. The acceptance ratios are shown here for each of the three replica exchange (REX) runs and the data are shown for Q_5 and Q_{15}. For Q_{30} and Q_{45}, the replica exchange simulations were initially carried out using a coarse, 11-window f_β schedule. However, the final PMFs were computed using the full 17-window f_β schedule shown in the legend for Figure 3.4. This was accomplished by collecting data from independent umbrella sampling runs (three per additional window) for windows that were not present in the coarse schedule. The decision to use the two-tier approach for the longer chains was based on the availability of CPU resources. The coarse f_β schedule used for replica exchange plus umbrella sampling simulations for Q_{30} and Q_{45} is as follows: [0.0, 0.1, 0.25, 0.3, 0.4, 0.5, 0.6, 0.75, 0.8, 0.9, 1.0].

The WHAM protocol has two components, namely iterative computation of relative free energies between adjacent windows and the construction of unbiased PMFs. Data from restrained simulations were analyzed using WHAM in three different ways. In the first two methods, both components of the WHAM protocol are used, with the only difference being the method used to compute error bars for the unbiased PMFs. The first method reports error bars as standard errors and the second method uses a bootstrap procedure to estimate error bars. For the third method, the relative free energies between adjacent windows were computed using thermodynamic integration (TI) and the unbiased PMF was computed using WHAM.

3.2.3 Calculation of PMFs for monomeric polyglutamine using WHAM with standard errors

As noted in the materials and methods section, we collected data from umbrella sampling simulations for each target f_β^0 value. For Q_5 and Q_{15}, we used umbrella sampling combined with replica exchange based on the 17-window f_β schedule shown in the main text. Data were collected from three independent replica exchange plus umbrella sampling simulations. For Q_{30} and Q_{45}, we combined umbrella sampling with replica exchange using a coarser 11-window f_β schedule and we carried out three independent simulations as well. The target f_β^0 values for each window in the coarse schedule were set to be: [0.0, 0.1, 0.25, 0.3, 0.4, 0.5, 0.6, 0.75, 0.8, 0.9, 1.0]. We subsequently added data from three independent simulations for each of the windows that were missing in the coarse schedule and were present in the finer schedule. In this way, we ended up with three independent f_β histograms (one from each run) for each of the seventeen windows shown in the Materials and Methods section. Each of three sets of 17 independent histograms were used in the WHAM analysis to construct three independent, unbiased PMFs. Panel A in Figure 3.3 shows the average PMF constructed from the three independent estimates.

The error bars in this figure are standard errors across the three independent estimates. The naïve standard errors represent the upper limit on the errors in our WHAM estimates of the PMFs.

3.2.4 Calculation of PMFs for monomeric polyglutamine using WHAM with bootstrap errors

For each window (with 17 windows in total), we have three independent simulations and for each of these simulations we recorded the f_β value every 5,000 MC steps. There were 4×10^7 production steps for Q_5,Q_{15}, and Q_{30}; therefore, we collected 8,000 f_β values for each window and for each independent simulation. For Q_{45} each run consisted of 6×10^7production steps, and we collected 12,000 f_β values for this chain. The f_β values samples were pooled across all independent simulations for each window. Thus, for each window characterized by a target f_β^0 value, the dataset contained 24,000 f_β samples for Q_5, Q_{15}, and Q_{30} and 36,000 f_β samples for Q_{45}. From this pooled dataset, we selected 800 f_β samples at random with replacement. This procedure was carried out for each window. These f_β values were binned to create an f_β histogram for each of the 17 windows with target f_β^0 values. Each histogram was created using 100 bins with bin widths of 0.1. This yielded a histogram for each of the 17 f_β windows. The standard WHAM analysis was applied to these constructed histograms to generate a PMF. This protocol of generating 17 sets of histograms using random sampling with replacement combined with WHAM analysis was repeated 200 times to generate 200 independent PMFs. Panel B in Figure 3.3 shows the average PMF (averaged over 200 trials) and bootstrap error, which is simply the standard error computed over the 200 unbiased PMFs.

3.2.5 Calculation of PMFs for monomeric polyglutamine using TI-WHAM with standard errors

For each independent replica exchange plus umbrella sampling simulation, we calculated $\partial E_i/\partial f_\beta^0{}_i$. This is the partial derivative of the system energy (including the restraint potential) for window i with respect to f_β^0. These statistics were gathered every thousand steps during the production run (4×10^7 for Q_5,Q_{15},Q_{30} and 6×10^7 for Q_{45}). Values for $\langle \partial E_i/\partial f_{\beta i} \rangle$ are numerically integrated using a cumulative trapezoidal integrator. This yields estimates for the free energy differences between all pairs of adjacent windows. These relative free energies were then fed into the WHAM analysis to combine f_β histograms for each window into a single unbiased probability distribution. The use of the TI step circumvents the iterative step in WHAM and allows us to analyze the PMF that results from an independent method for assessing relative free energies between windows. For the TI-WHAM procedure, we used data based on the coarse (as opposed to fine) f_β schedule. This allowed us to query the dependence of the reconstructed PMFs on the coarseness of the f_β schedule (in addition to the circumvention of the first, iterative step in WHAM). For each chain length, the TI-WHAM procedure was applied to each of the three independent datasets. The PMF shown in Panel C of Figure 3.3 is an average over the three runs and the error bars are standard errors.

For polyglutamine dimer simulations, we combined MC simulations with thermal replica exchange. In two of the three sets of simulations, each chain was restrained to target f_β^0 values of = 0.75 and 1.0, respectively, while the third simulation set involved unrestrained molecules. For each chain length, we carried out three independent replica exchange runs. The following Kelvin temperature schedule was used for the replica exchange simulations: [298, 305, 315, 325, 335, 345, 355, 360, 370, 380, 390]. The temperature schedule was based on data for coil-to-globule transitions of unrestrained monomeric polyglutamine. As shown in Figure 3.7, temperature modulates the solvent quality. Temperatures in the range $T \leq 355$K correspond to the poor

solvent regime, where monomeric polyglutamine prefers globules; temperatures in the range 355 K < T < 420 K are in the transition region between globule and coil, and higher temperatures correspond to swollen coils. The theta point (T_θ) was found to be approximately 390K (see caption to Figure 3.7 in the supplementary material). We wish to quantify the spontaneity of intermolecular associations in the poor solvent regime. The overlap between coil and globule ensembles is small and decreases with increasing N. Therefore, we set the upper limit for the replica exchange temperature schedule to be T_θ, to ensure that the replicas were used judiciously. We quantified overlap statistics and acceptance ratios for swaps between adjacent thermal replicas to demonstrate the adequacy of the protocol, and details are in Figures 8 – 10.

3.2.6 Analysis of thermal replica exchange data for dimers of polyglutamine

Energy values are collected every 5,000 steps. From the production runs we obtained 7,900 energy values for Q_5, Q_{15}, and Q_{30} 12,000 samples for Q_{45}. We constructed histograms of energy values using placed 200 bins (with bin widths that vary with chain length). All error bars were standard errors computed across three independent runs.

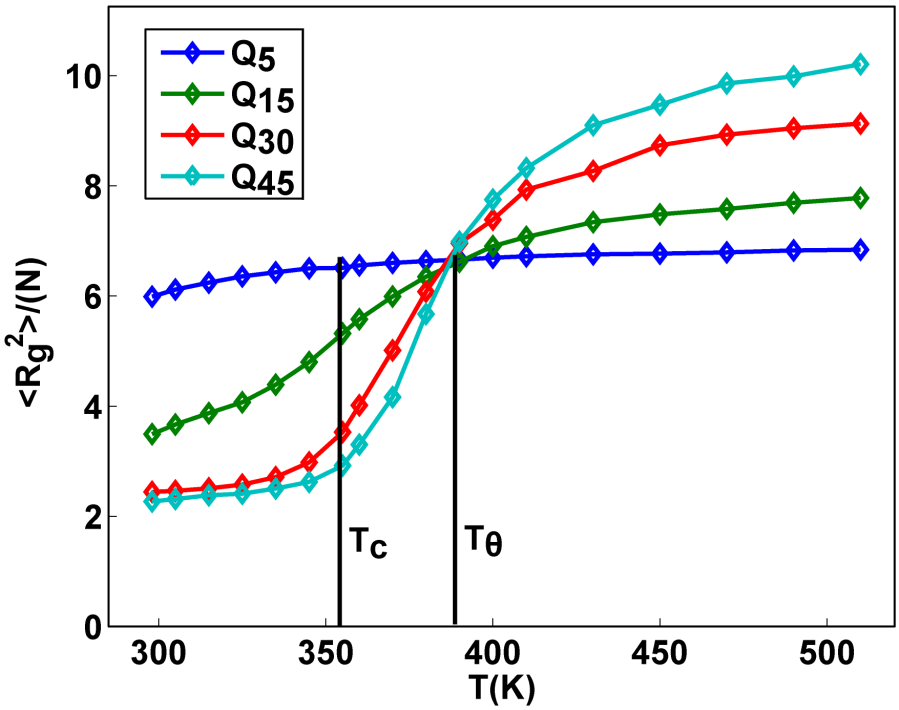


Figure 3.7: Coil-to-globule transition for monomeric polyglutamine. These data show the ensemble averaged mean square radius of gyration as a function of simulation temperature. Values on the ordinate are normalized by chain length because $\langle R_g^2 \rangle \propto N$ at the theta temperature, T_θ. Therefore, the plots for different chain lengths should coincide at $T=T_\theta$ for different chain lengths because polyglutamine molecules are homopolymers. This requirement is satisfied for $T=T_\theta \approx 390$K. The plots were made by analyzing data extracted from torsional space Metropolis Monte Carlo (MC) simulations using the forcefield and ABSINTH implicit solvation model described in the Materials and Methods section of the text. Details of the move sets used are provided in Table 3.1. Independent simulations were carried out for each temperature. For each temperature, the simulations involved 10^6 equilibration steps of MC followed by 4×10^7 steps of production. A complete analysis of the coil-to-globule transition for monomeric polyglutamine and an explanation for the shapes of these curves has been provided in previous work [13].

$T=T_C\approx355$K is the temperature below which collapsed states are favored for polyglutamine. Hence, for the forcefield used in this work, $T \leq T_C$ corresponds to poor solvent conditions, whereas the regime $T_C < T \leq T_\theta$ corresponds to the transition region.

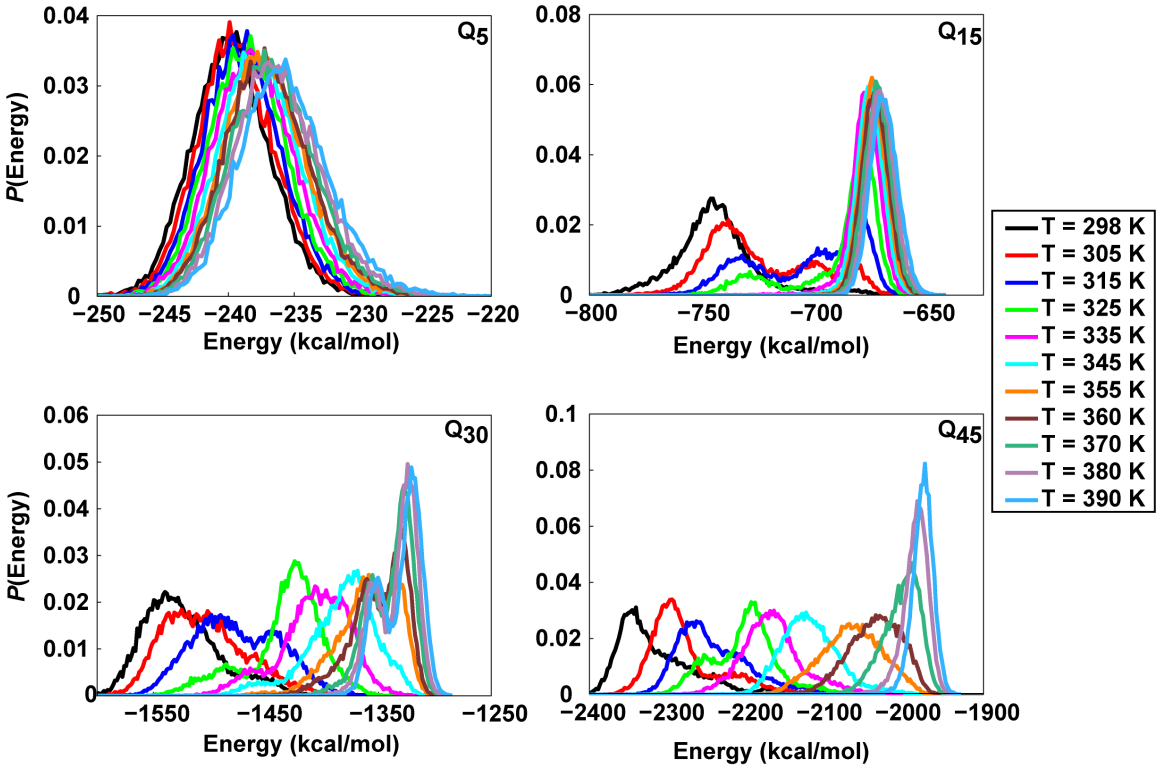


Figure 3.8: Temperature-dependent energy histograms for dimers of polyglutamine molecules where both chains are restrained to a target value of f_β^0=1. The histograms demonstrate the overlap between adjacent replicates and justify the temperature schedule used in all replica exchange simulations of homodimerization. The overlap between distal replicas decreases with increasing chain length and this is consistent with the differences in thermal stabilities for associated versus dissociated dimers for different chain lengths (longer chains form associated dimers over a wider temperature range). As noted in the text, we carried out three independent thermal replica exchange runs for each chain length and target f_β value. The figure

shows average histograms for each temperature. Figure 3.9 summarizes the overlap statistics that were calculated using the histograms shown in this Figure.

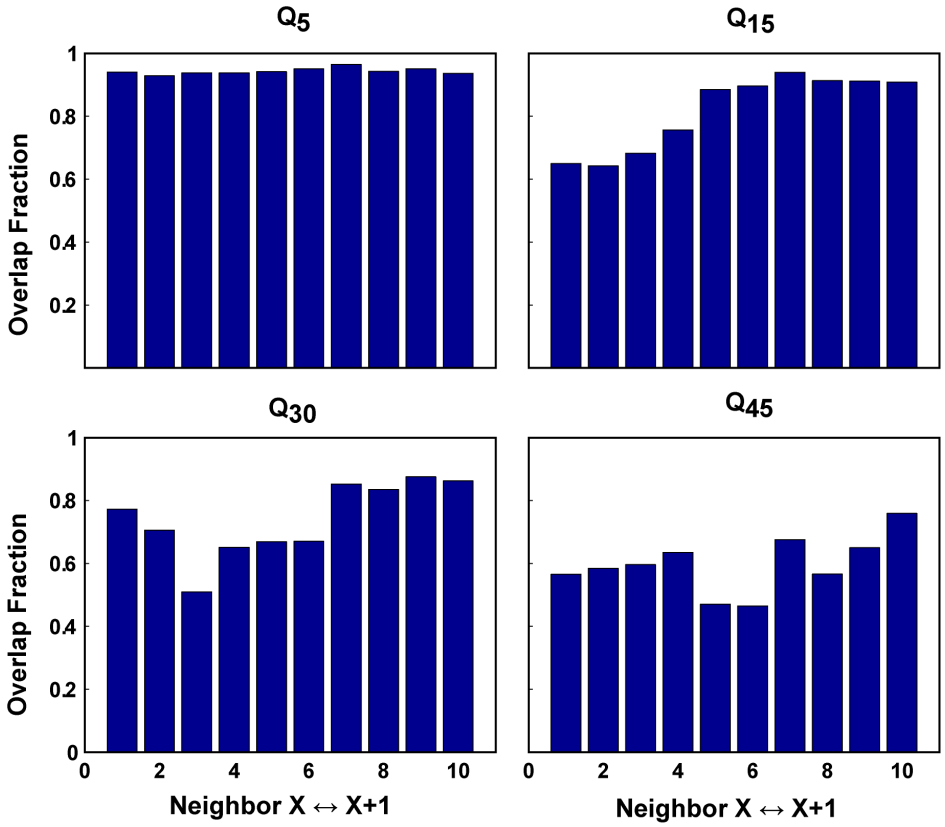


Figure 3.9: Quantification of the overlap between adjacent thermal replicas shown in Figure 3.8. Ten bars are shown for overlap statistics computed between eleven sets of adjacent windows. The temperature schedule is presented in the methods section. Computation of the overlap statistics is analogous to the method described in the caption for Figure 3.3. The overlap statistics appear to justify the adequacy of the temperature schedule used in the replica exchange runs.

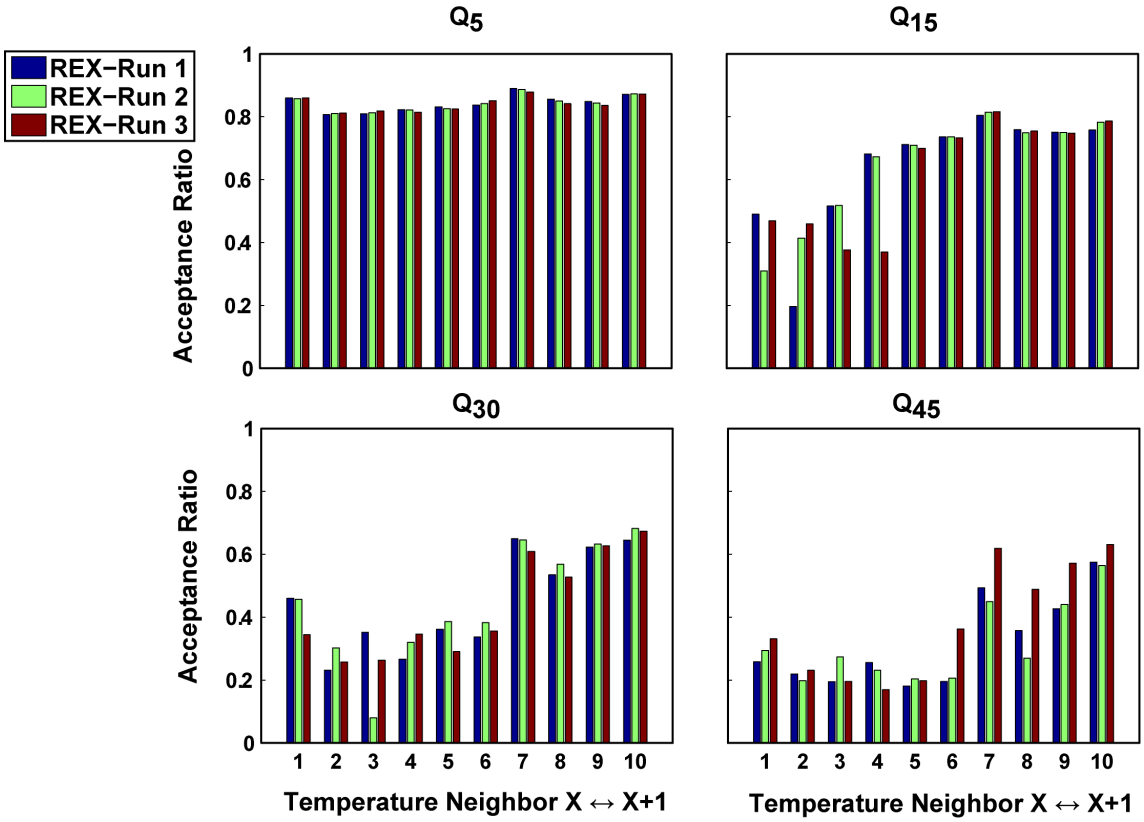


Figure 3.10: Plot of acceptance ratios for swaps between nearest-neighbor thermal replicas in simulations of homodimerization. Data are shown for the case where both chains are restrained to a target value of f_β^0=1. The acceptance ratios are shown here for each of the three replica exchange (REX) runs that were carried out for each peptide. Even for the longer chains, the smallest acceptance ratios are greater than 0.2 (on average) indicating that the sampling quality is not compromised.

3.3 Conformations with high β-content are thermodynamically unfavorable for monomeric polyglutamine

Panel A of Figure 3.11 show free energy profiles for monomeric polyglutamine at four chain lengths. Panels B and C show the same free energy profiles, except that the profile in panel B shows the bootstrap errors, whereas panel C shows the free energy profiles constructed using

the TI-WHAM procedure. The free energy profiles along f_β are smooth and possess broad minima for $N \geq 15$. For Q_5, the PMF is essentially flat because the peptide is short and cannot collapse. For longer chains, the minima are located at values of $f_\beta \approx 0.3$, which is the value preferred by unrestrained polyglutamine chains. The Q_{15} chain possesses a tendency for forming α-helical segments [13]. This helix propensity is the reason for the flatness of the free energy profile for values of f_β less than 0.3 for this peptide. Decreased α-helical propensities for longer chain lengths lead to an increase in the free energy penalty for $f_\beta < 0.3$.

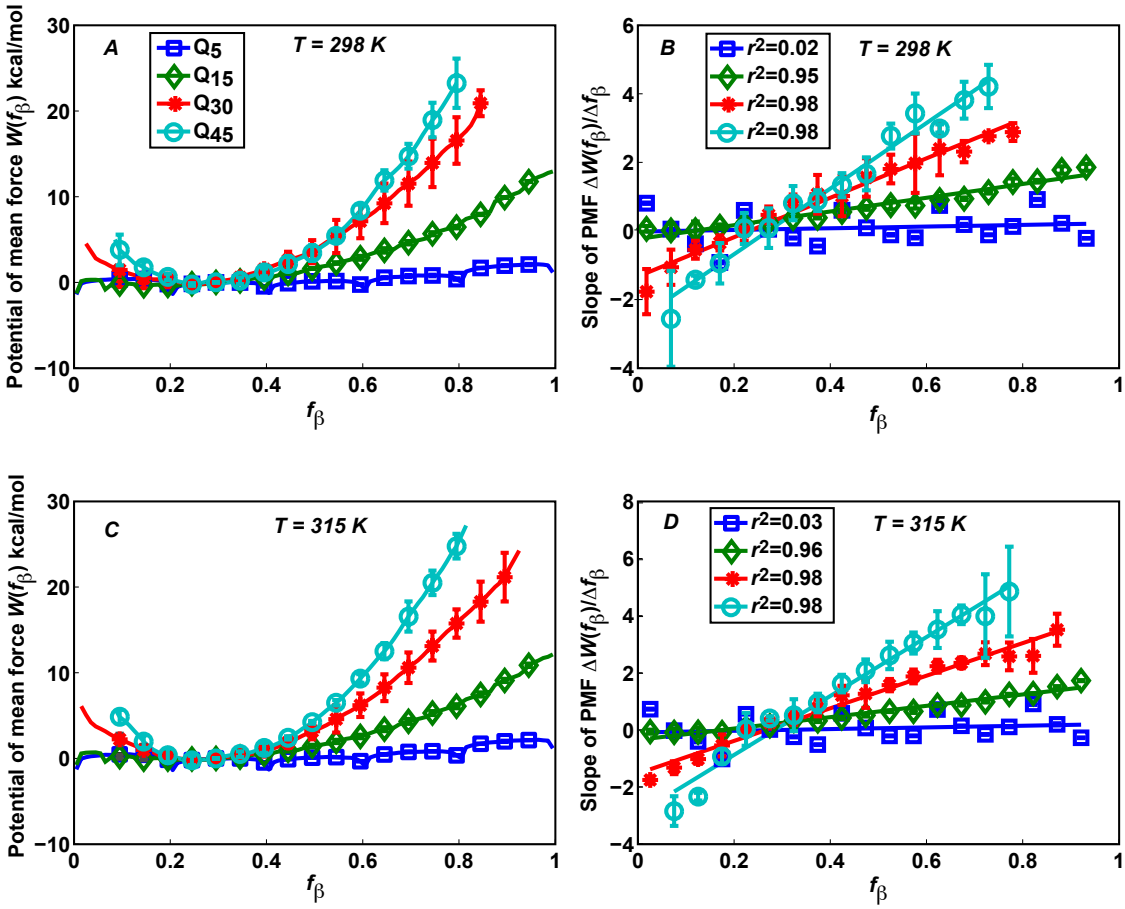


Figure 3.11: Potentials of mean force (PMFs) for monomeric polyglutamine chains of different lengths. The profiles are plotted as a function of f_β, which is a reaction coordinate that measures β-content. Panel A shows the PMFs with standard errors, panel B with bootstrap

errors, and panel C shows the PMFs that result from using TI-WHAM with a coarse f_β schedule. Details are described in the supplementary material. The lines in panel D are lines of best fit to the data for the derivatives of the PMFs shown in panel A. The insets show the Pearson correlation coefficients – r^2 – that diagnose the strength and direction of the hypothesis that the PMF derivatives (forces) are linear. The slopes and intercepts for the lines of best fit are as follows: Q_5 (0.25, -0.05), Q_{15} (2.00, -0.27), Q_{30} (5.74, -1.38), Q_{45} (9.5, -2.65). Slopes have units of kcal / mol-f_β and intercepts have units of kcal / mol.

The free energy profiles do not show evidence for distinct local minima. This is confirmed by analyzing the derivatives of each of the PMFs. In the harmonic limit with a single minimum, a mean force profile will be a straight line and this is what we observe as shown in panel D of Figure 3.3. Therefore, monomeric polyglutamine does not access metastable conformations with high f_β values. Bhattacharya *et al.* [15] estimated the nucleation free energy for Q_{47} to be roughly 12 kcal / mol. β-rich structures will have f_β values between 0.6 and 0.7 (Figure 3.2 and Figure 3.4). Free energy penalties associated with accessing such structures are in the range of 10-15 kcal/mol for Q_{45} (Panel A in Figure 3.11). This is in the range of the estimate obtained by Bhattacharya *et al.*

Bhattacharya *et al*. also proposed that aggregation rates increase with increasing chain length because free energy penalties associated with forming β-rich structures decrease with increasing chain length [12, 15]. The calculated free energy profiles do not support this hypothesis. The free energy penalty for forming structures with high f_β values increases with increasing chain length. However, this increase in free energy penalty levels off for Q_{45} because the PMFs for Q_{30} and Q_{45} resemble each other – more so than those for Q_{30} and Q_{15}. We previously showed that the stability of non-specifically collapsed states increases with increasing chain length [13]. Here,

we show that the likelihood of populating structures with high f_β values decreases with increasing chain length. The former observation is consistent with the physics of flexible polymers in a poor solvent, whereas the latter observation is congruent with the idea that energetic frustration [30] makes it difficult for flexible, homopolymeric polyglutamine to adopt well-ordered three-dimensional structures in their monomeric forms.

3.4 Restraining monomeric polyglutamine to high f_β values provides access to ordered β-sheet conformations

Fractional DSSP-E scores [28] based on analysis of hydrogen-bonding patterns are a complementary measure of secondary structure. Figure 3.12 plots data for monomeric polyglutamine with fractional DSSP-E scores along the ordinate and f_β values along the abscissa. Data were extracted from the simulations that were performed in the presence of restraints on f_β. We do not observe any structures for which the f_β value is low when the fractional DSSP-E score is high. Structures with characteristic β-sheet hydrogen bond patterns have high fractional DSSP-E scores as well as high f_β scores. However, for moderate to high f_β values there is a substantial spread in the observed fractional DSSP-E scores, implying that polyglutamine samples disordered conformations lacking regular backbone-backbone hydrogen bonds even when f_β is restrained to a high value.

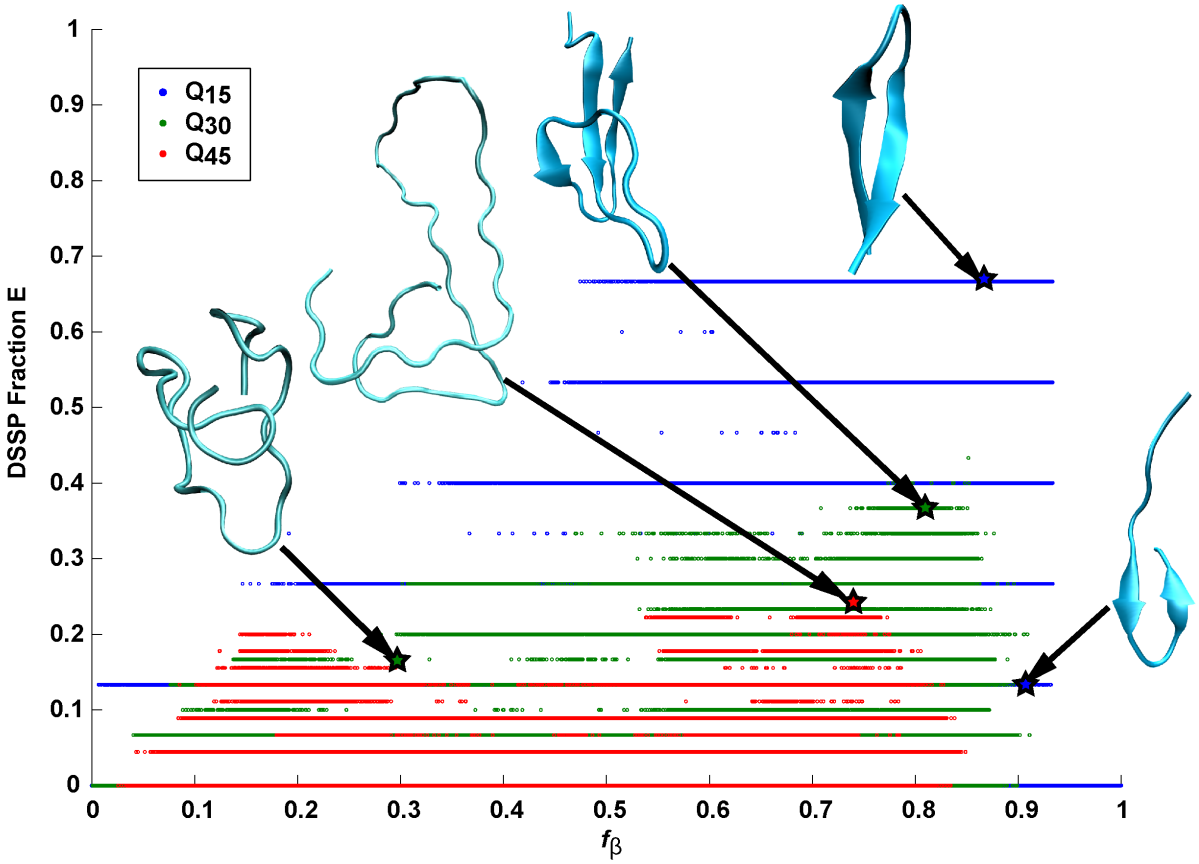


Figure 3.12: Scatter plot of all recorded snapshots in all simulations for Q_{15}, Q_{30}, and Q_{45} correlating the fractional β-content according to DSSP with the values for f_β at 298 K. Dots of different colors correspond to chains of different length. Representative points are marked using stars and the corresponding structures are shown in cartoon representation. Graphics were generated using VMD [31]. Note that the fractional β-content according to DSSP is an inherently discrete quantity for chains of finite length. Q_5 is not shown since the chain is too short to have non-zero DSSP-E scores.

3.5 Persistence of disorder in the presence of restraints results from a diverse registry of intramolecular hydrogen bonds

We asked if disordered structures contain large numbers of hydrogen bonds, which are either unsatisfied or assumed to be satisfied by the solvent? In Figure 3.13, we plot the average number of intramolecular hydrogen bonds per backbone and sidechain oxygen (acceptor) atom in monomeric polyglutamine as a function of chain length. All four types of hydrogen bonds including backbone acceptor to backbone donor, sidechain acceptor to sidechain donor, backbone acceptor to sidechain donor, and sidechain acceptor to backbone donor, make roughly equivalent contributions to the total hydrogen bond inventory. No obvious preference for specific intramolecular contacts is seen, even in the presence of restraints. Summing up around the acceptor sites, the average number of hydrogen bonds per backbone / sidechain acceptor is less than unity, indicating that a substantial fraction of hydrogen bonds are satisfied by the solvent. The average number of intramolecular hydrogen bonds per acceptor is generally larger for the unrestrained case. Restraints to high values of f_β increase the number of solvent-exposed acceptors leading to increased interfaces with the surrounding solvent. Despite the restraints, the peptides prefer the full spectrum of intramolecular and chain-solvent contacts.

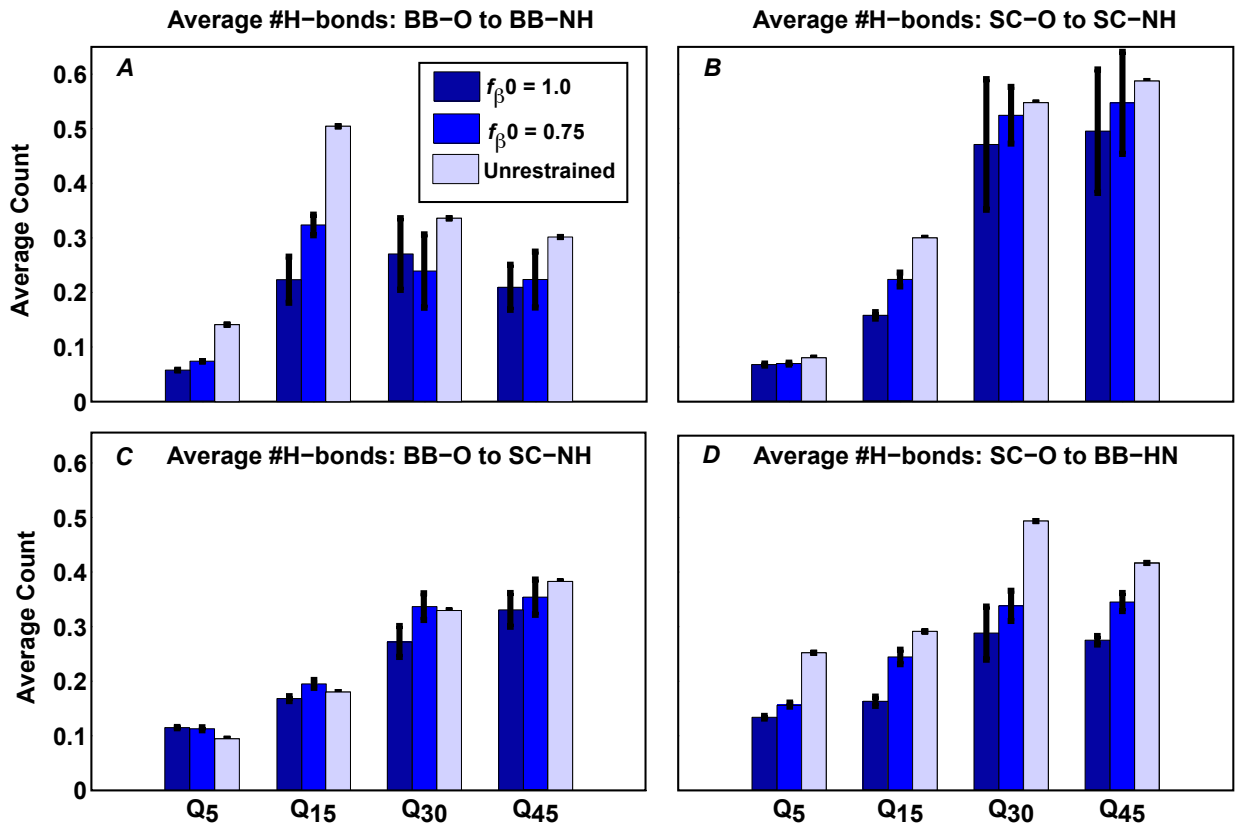


Figure 3.13: **Average number of hydrogen bonds per acceptor oxygen atoms.** Data are shown for hydrogen bonds around the backbone (Panels A and C) and sidechain (Panels B and D) acceptor oxygen atoms, respectively. BB denotes backbone and SC denotes sidechains; Values are shown for T = 298 K and three different chain lengths (Q_{15}, Q_{30}, Q_{45}) and restraint values. Hydrogen bonds were determined using the general definition introduced by Kabsch and Sander *(28)*.

3.6 Dimerization of polyglutamine remains spontaneous in the presence of restraints

We asked if the spontaneity of homodimerization is altered when each chain is restrained to adopt high values of f_β. We simulated dimerization of polyglutamine for chains of length N=5, 15, 30, 45 as a function of temperature and two different values for the target restraint, f_β^0. If the

homogeneous nucleation model applies for polyglutamine aggregation, then associations involving restrained chains should be stronger than the associations involving unrestrained chains. To quantify associativity we computed a temperature-dependent excess interaction coefficient $B_{22}(T)$:

$$B_{22}\left(T\right) = \frac{\int_{R=0}^{R=d_{droplet}} \left[F_{T=T_\theta}\left(R\right) - F_T\left(R\right) \right] R^2 dR}{\int_{R=0}^{R=d_{droplet}} F_{T=T_\theta}\left(R\right) R^2 dR} \tag{3.4}$$

Here, $F_T(R)$ is the cumulative distribution function for the intermolecular distance at temperature T; $d_{droplet}$ = 400Å is the diameter of the simulation system. When spontaneous associations are favored vis-à-vis T_θ, $B_{22}(T)$ is negative. If chains interact the way they would at T_θ, then $B_{22}(T)$ is zero.

Figure 3.14 plots $B_{22}(T)$ as a function of T for different chain lengths in the presence and absence of restraints on f_β. The associativity of chains is length and temperature-dependent, but it does not vary with the presence or absence of conformational restraints. Panel A in Figure 3.14 shows that the short Q_5 peptide remains non-associative across the entire simulated temperature range. For Q_{15}, the differences in associativity with and without restraints are statistically insignificant. At the lowest temperatures, B_{22} is negative for restrained as well as unrestrained chains. It increases monotonically with increasing temperature and reaches zero between 320 and 340 K. This temperature range represents the stability limit for the associated dimer. Panels B and C of Figure 3.14 show the same data for Q_{30} and for Q_{45}. As chain lengths increase, the stability limit for associated dimers shifts to higher temperatures. Within the statistical accuracy of the data, chains with restraints are as associative (not more associative) as unrestrained chains.

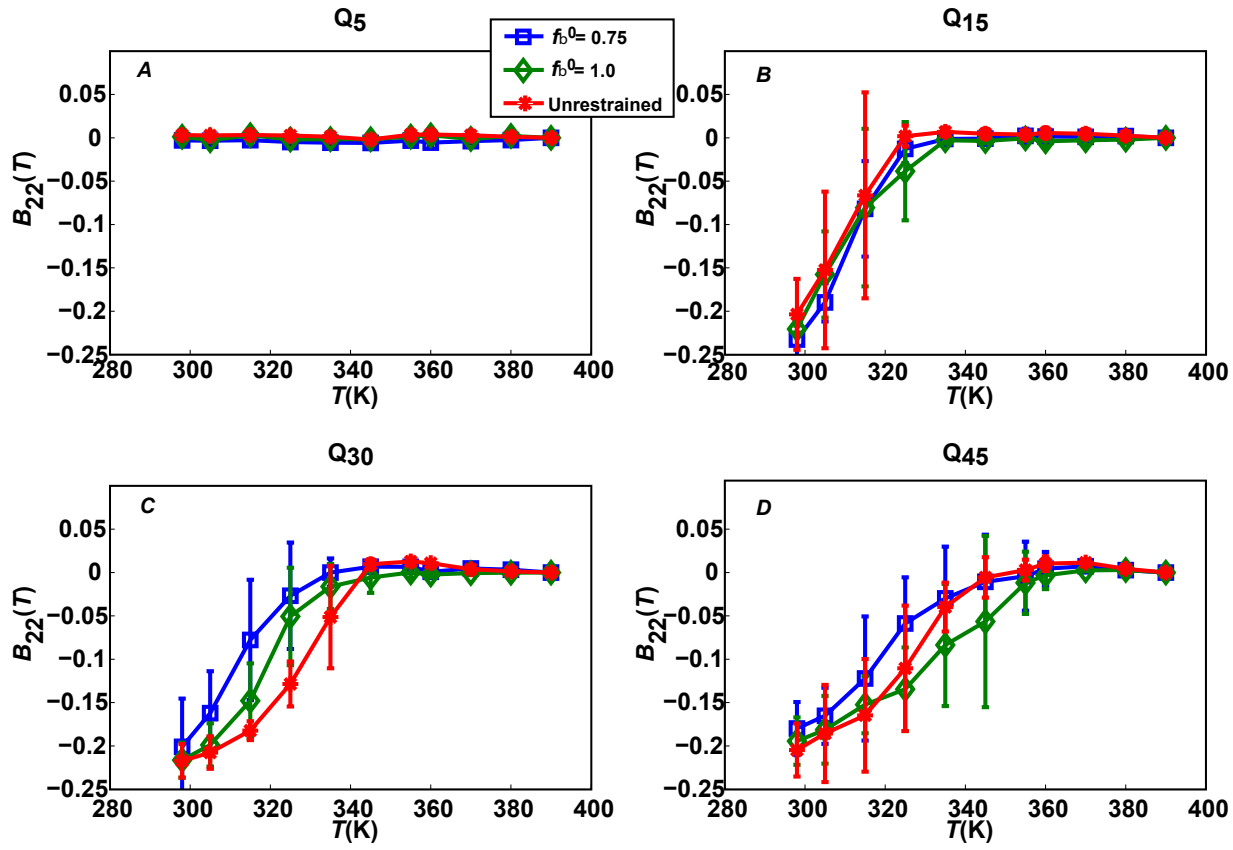


Figure 3.14: **Plots of $B_{22}(T)$ as a function of temperature.** Each panel shows $B_{22}(T)$ extracted from simulations with unrestrained chains and simulations where each chain in the simulation has a target restraint of $f_\beta^0 = 0.75$ or $f_\beta^0 = 1.0$.

For high values of f_β, the surface-to-volume ratio increases with respect to the unrestrained chains as shown in Figure 3.15. This does not translate into increased associativity with respect to the unrestrained chains. Instead, the actual poorness of the solvent (the value of T) remains the main determinant of the spontaneity of intermolecular associations. This observation is confounding since solute-solvent interfaces increase for monomeric polyglutamine when restraints toward high f_β are imposed. We quantified the solvent-solute interface for monomeric polyglutamine by asking if interfacial energies were insensitive to the presence or absence of conformational restraints. We decomposed the system energies (excluding the

restraint potential) into bulk (volumetric) and surface contributions [32] by fitting the system energies to the functional form shown in Equation 3.5. Here, C_1 and C_2 measure the effective strengths of volumetric self-interactions and surface terms, respectively.

$$\frac{\left\langle U_{\text{total}} - U_{\text{restr}} \right\rangle}{N} = C_1\left(T, f_\beta^0\right) + C_2\left(T, f_\beta^0\right) \cdot N^{-1/3} \tag{3.5}$$

N is the chain length, U_{total} is the system energy, and U_{restr} is the restraint potential energy. Figure 3.16 plots C_1 and C_2 as a function of temperature for different values of f_β^0. When compared to the unrestrained simulations, the C_1 term in the collapse regime is less favorable in the presence of restraints indicating the high free energy of structures with high values of f_β. Additionally, for a given temperature, values for C_2 are more positive because larger interfaces with the surrounding solvent are more unfavorable than the smaller solute-solvent interfaces made by unrestrained chains. As $T \rightarrow T_\theta$, C_2 approaches zero indicating that the interface is indifferent with respect to interactions with either solvent or the chain. This is Flory's definition of the theta-state. For temperatures greater than $T_\theta \approx 390$ K, C_1 converges to a value of ca. –20 kcal/mol. This is the approximate mean-field energy for a fully solvated glutamine residue within the ABSINTH Hamiltonian [14], and is the expected limiting value for a chain preferring chain-solvent interactions to chain-chain interactions. Therefore, in the presence of restraints toward high target values of f_β the chains swell and form energetically unfavorable interfaces with the surrounding solvent. However, these features do not translate into differences in B_{22} vis-à-vis the unrestrained chains.

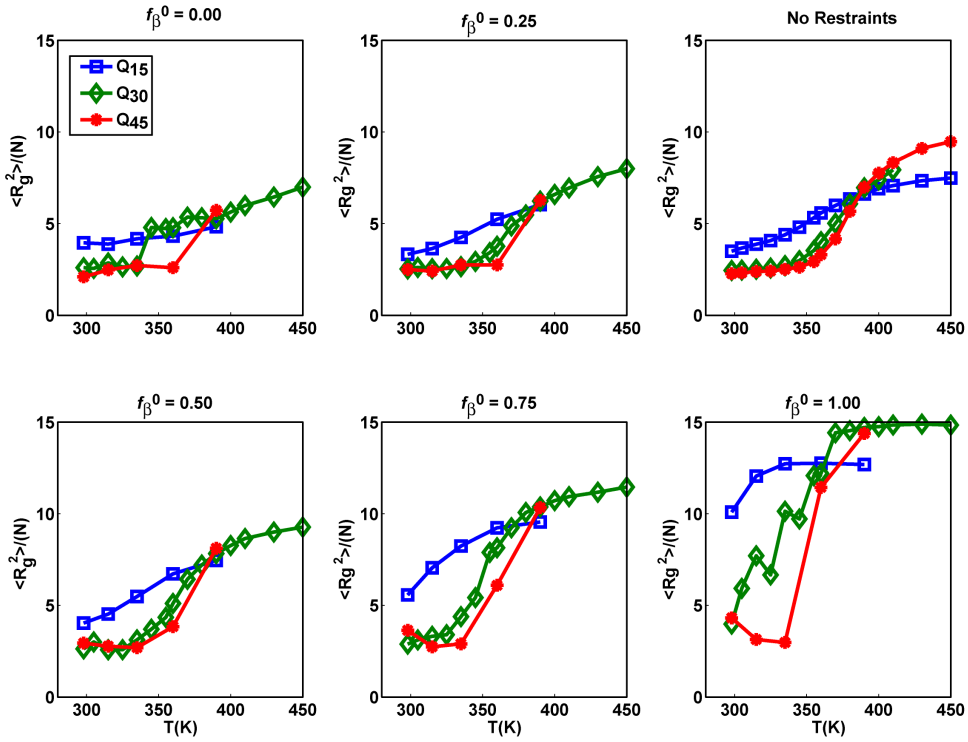


Figure 3.15: Quantitative analysis of the effects of conformational restraints on coil-to-globule transitions for monomeric polyglutamine of different lengths. Two trends are apparent in the data: First, the coil-to-globule transitions becomes sharper vis-à-vis the unrestrained case for target values of f_β greater than 0.5, whereas the opposite is true for $f_\beta \leq$ 0.25. Second, the normalized value of $\langle R_g^2 \rangle$ remains in the vicinity of the value for the unrestrained chain for temperatures that are in the globule regime; conversely, for temperatures in the coil regime the normalized $\langle R_g^2 \rangle$ values are significantly larger than unrestrained values if $f_\beta \geq 0.75$, whereas the opposite is true for $f_\beta \leq 0.25$.

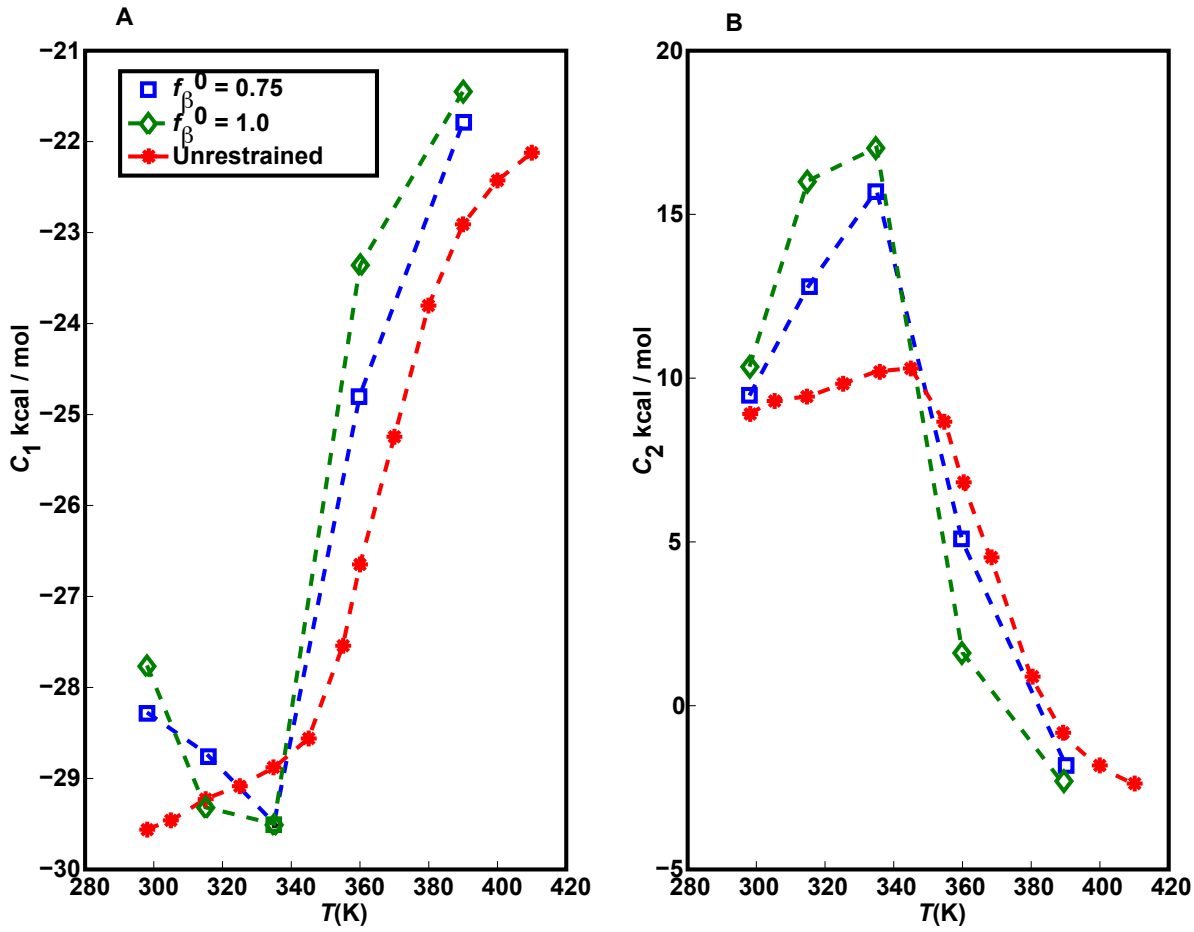


Figure 3.16: Energy density C_1 (Panel A) and surface energy term C_2 (Panel B) for monomeric polyglutamine. Data were obtained for unrestrained polyglutamine and two other simulations with restraints on f_β. The quality of the fits underlying these data cannot be accurately assessed since they are fits to data from only three chain lengths (because we exclude Q_5 excluded from analysis, since it is too short for a volumetric term to contribute).

3.7 Restraints toward high values of f_β promote the formation of canonical β-sheets

Figure 3.17 shows the average fractional DSSP-E scores observed in monomer and dimer simulations with f_β^0=1 compared to the data obtained from simulations with unrestrained chains.

When the free energy penalty for sampling high f_β values is pre-paid, the resultant increase in solute-solvent interface at the monomer level results in increased β-sheet formation through a coupling between inter- and intramolecular interactions. This difference is suggestive of intermolecular interfaces being responsible for promoting β-sheet content. To provide a visual perspective of increased β-sheet content in dimers formed with restrained chains, Figure 3.18 shows a scatter plot in the space of the two parameters, f_β and fractional DSSP-E scores. Figure 3.19 shows a bar plot of intermolecular hydrogen bonds to quantify the interactions in intermolecular interfaces. Intermolecular hydrogen bonds predominantly involve glutamine sidechains. There is a significant enhancement in intermolecular backbone-backbone hydrogen bonds for Q_{15} and Q_{30} when $f_\beta^0 = 1.0$ and is congruent with data in Figure 3.17. Interestingly, the per-molecule PMF obtained via WHAM for two Q_{30} molecules in the simulation system is virtually identical to the one obtained for the monomer (data not shown). Hence, we conjecture that larger oligomers must form to facilitate spontaneous conversion of individual chains to β-rich conformations.

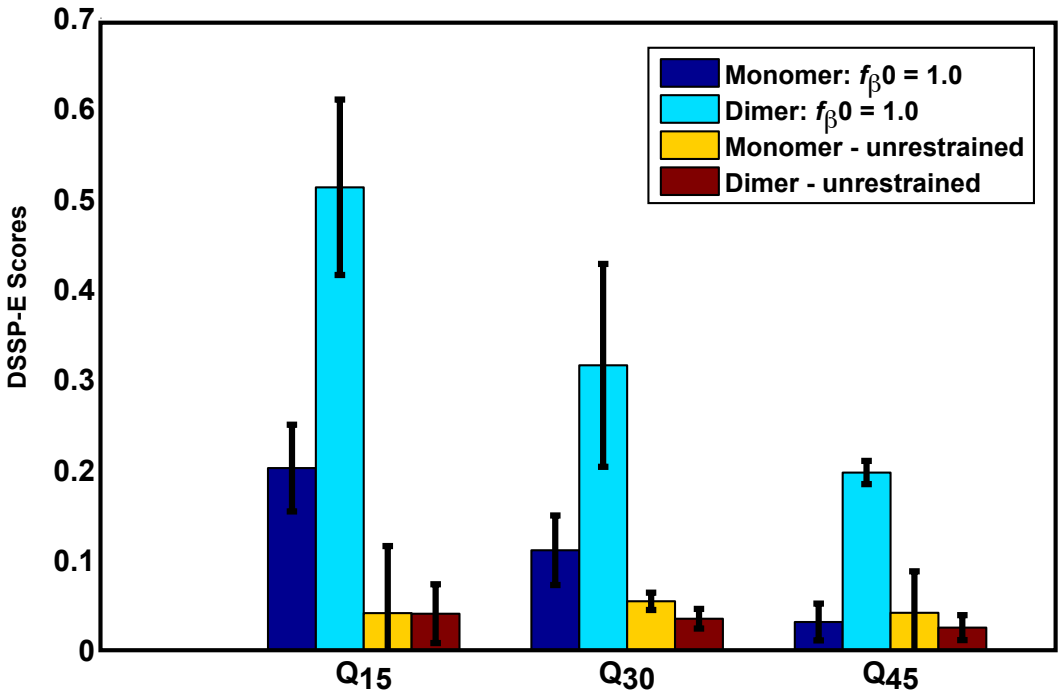


Figure 3.17: **Bar plots comparing the average fractional DSSP-E scores in simulations of monomeric polyglutamine to simulations of dimeric polyglutamine at 298 K.** Data are shown for three chain lengths using data from simulations with unrestrained chains as well as data from simulations for f_β^0=1.0.

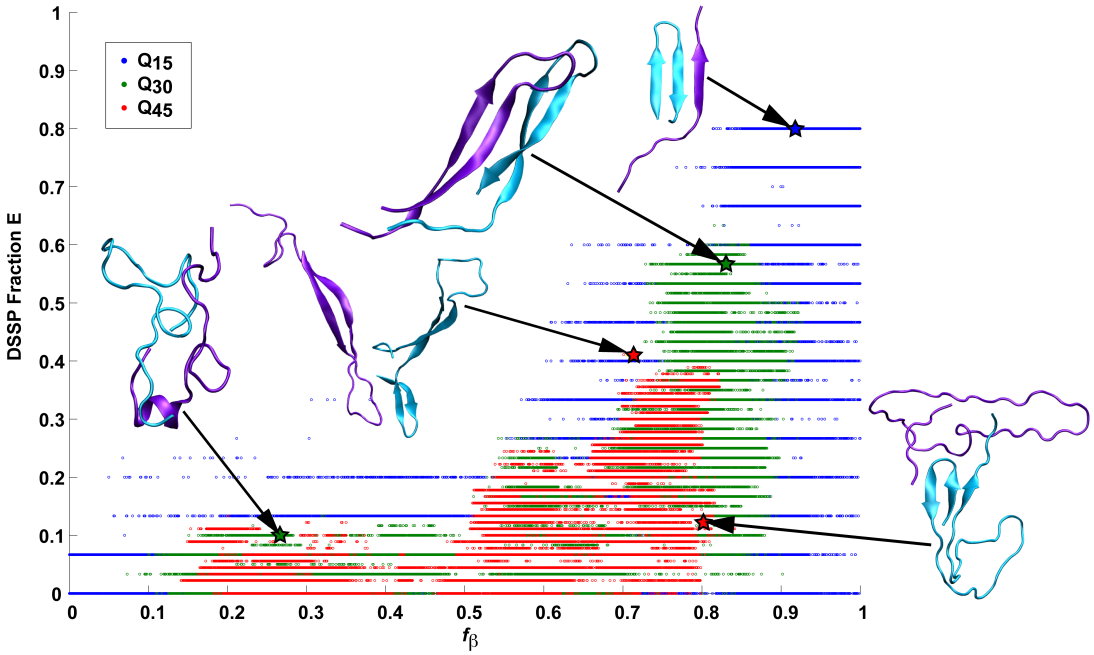


Figure 3.18: Scatter plot of all recorded snapshots in all dimer simulations for Q_{15}, Q_{30}, and Q_{45} correlating the fractional β-content according to DSSP with the values for f_β at 298 K.

This plot is analogous to the one shown in Figure 3.12.

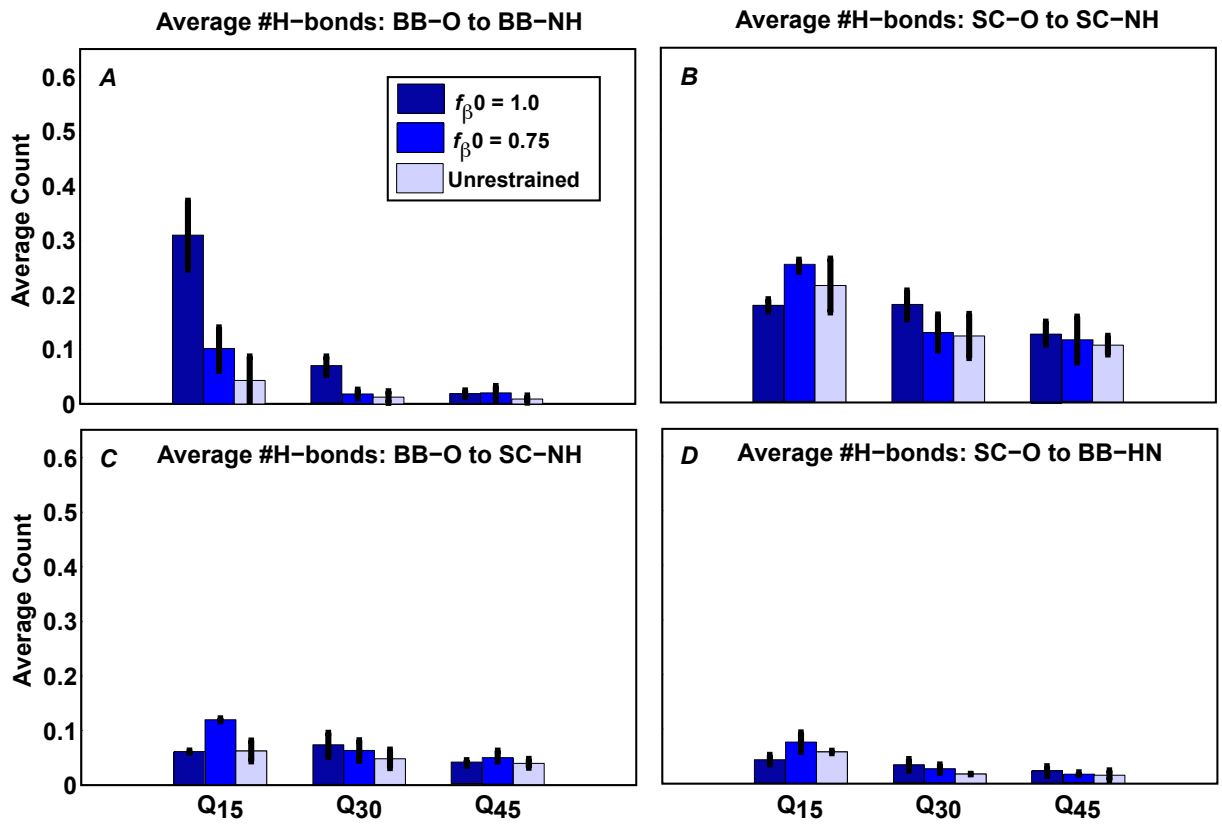


Figure 3.19: Average number of hydrogen bonds per acceptor oxygen atoms. This plot is similar to Figure 3.13 except that the hydrogen bond statistics are shown for simulations with two molecules. Only intermolecular hydrogen bonds are shown. Q_5 is excluded from this plot, since no intermolecular hydrogen bonds are detected in these simulations.

3.8 Addendum: improvements to the f_β restraining potential

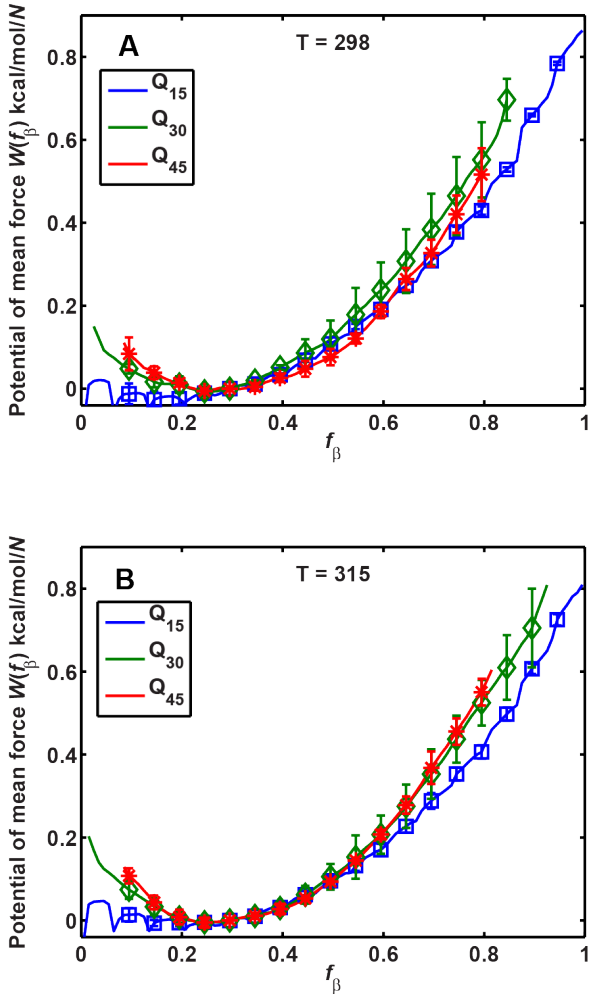


Figure 3.20: Potentials of mean force (PMFs) for monomeric polyglutamine chains normalized by chain length. The PMFs shown in Figure 3.11, panels A and C, were divided by their respective chain lengths and this yields the free energy penalty per reside of occupying the β-basin. After normalization, all cures are nearly within error of each other.

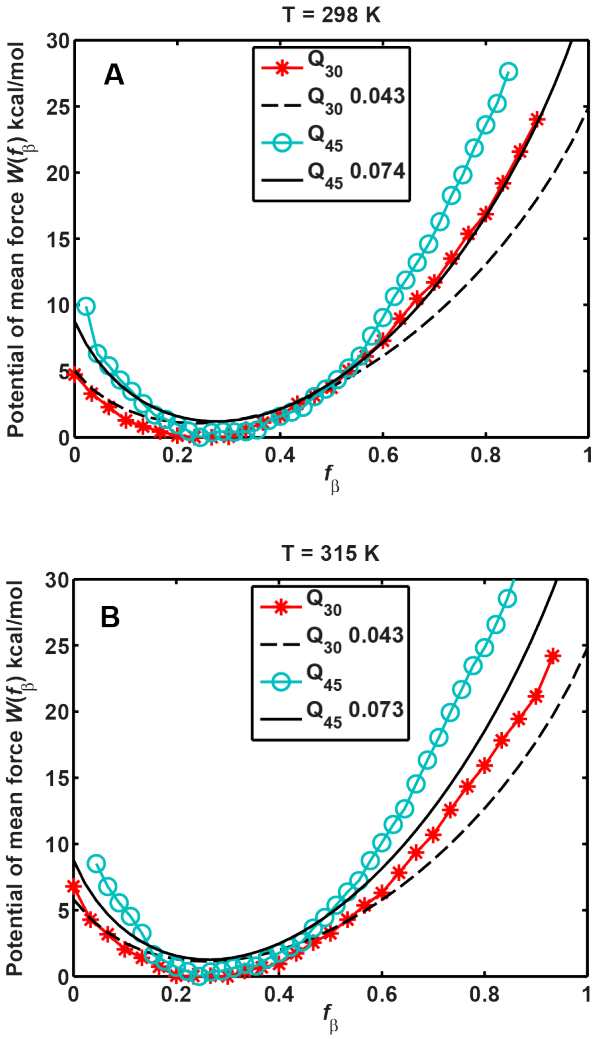


Figure 3.21: Potentials of mean force (PMFs) for monomeric polyglutamine chains fit to a binomial probability distribution. The PMFs for Q_{30} and Q_{45} in Figure 3.11 panels A and C are shown as red (*) and cyan (O) curves, respectively. The black lines are fits for each chain length to a binomial probability distribution function P_b and this is defined as the probability of observing f_β given a chain length N where the probability of occupying the β-basin per residue is $P_{f\beta r}$. The minimum Kullback–Leibler divergence was computed between $P(f_\beta)$ obtained from WHAM and P_b by adjusting the free parameter $P_{f\beta r}$. This minimum is shown in the figure legend. These probability distributions are expressed as PMFs by computing $-RT\log(P)$. Kullback–

Leibler divergence was computed from $\sum_i \log(\frac{P_{f\beta}(i)}{P_b(i)})\, P_{f\beta}(i)$ where i visits each f_β bin of size 0.01.

After publication of this work there was an important insight regarding $W(f_\beta)$ in Figure 3.11, namely, $W(f_\beta)$ measures the entropic penalty associated with placing residues in the β-basin. Figure 3.20 makes this point by showing the chain-length normalized PMFs collapse on one another. Figure 3.21 shows that most of the information contained in the PMFs can be represented by a binomial, which quantifies the probability distribution of f_β where the independent probability for each residue to occupy the β-basin is fixed. Conformance to the binomial distribution implies there is no cooperativity between residues in forming β-strands and disorder dominates. This highlights the weakness of this restraining potential it that it allows for highly heterogeneous conformations devoid of ordered canonical β-strands to satisfy the potential for high values of f_β. This increased heterogeneity means it is more difficult for the underlying potential to sample ordered conformations for high f_β which defeats the primary goal of a structural restraint potential.

A new restraint potential using DSSP-E as a reaction coordinate was developed and tested but found to be insufficient as it imposed serious sampling challenges that were not immediately addressable. Future work involves computing the stabilities of highly ordered and highly disordered conformations on the manifold of high f_β. These results will guide an improved sampling protocol whereby a bias is inserted for more ordered, canonical β-strand conformations and away from disordered conformations. This improved dual-restraint technique will be used to re-asses the conclusions from this work.

3.9 Discussion

Monomeric polyglutamine encompassing the threshold length for Huntington's disease (HD) does not adopt conformations rich in β-content. Such conformations have been proposed as putative nuclei for polyglutamine aggregation [12]. Our estimate of the free energy penalty associated with forming structures with high β-content is in the range of estimates proposed in the literature [15]. However, these penalties increase with increasing chain length and may plateau past a certain chain length. Our observations contradict the expectations of Bhattacharya et al. [15] who proposed that the free energy penalties in question should decrease with increasing chain length if the stability of the monomeric nucleus is the source of the increased rate of aggregation with increasing chain length. In our calculations, coil-to-globule transitions for monomeric polyglutamine (Figure 3.7) as well as the spontaneities of intermolecular association show clear dependence on chain length. In a poor solvent, longer chains form disordered globules of increased stability. The more stable the globules, the more favorable the dimers [13] and the harder it is to dissociate the dimers. While conformational restraints toward high f_β values lead to increases in solute-solvent interfaces and increased β-content, these do not translate into increased spontaneity for intermolecular associations. This observation points to poorness of the solvent, as opposed to the associativities ascribed to specific structures, as the invariant and generic driving force for promoting aggregation of homopolymers. Even though high β-content remains just as thermodynamically unfavorable at the dimer level as at the monomer level, the formation of intermolecular interfaces between restrained chains promotes the formation of canonical backbone-driven β-sheet structures. Other approaches to simulating polyglutamine conformational equilibria and aggregation have been published [33, 34]. Clearly, comparative

calculations using different models are needed to test the model dependence of our predictions and this is part of ongoing work.

Recent experiments [35] suggest that β-sheet formation is a feature of large aggregates. It was also shown that large non-specific aggregates form prior to nucleated formation of ordered fibrils. These observations are consistent with extrapolations from our calculations, specifically the data shown in Figure 3.19. In Figure 3.1, we summarize conceivable routes from the ensemble of disordered globules to the high molecular weight, ordered fibrillar aggregates. The homogeneous nucleation model, with a monomeric nucleus is highlighted in gray. An alternative proposal, consistent with our results and recent experimental data of Lee et al. [35] is follows: Disordered globules reversibly associate to form a broad distribution of disordered oligomers. These oligomers can be described as being "molten" [18] or "liquid-like" [16] and may be large enough to be referred to as "mesoglobules" [36]. Molten oligomers are peptide-rich microphases, and chains in the interior are solvated by other chains. These peptide-peptide interfaces should be more favorable than peptide-solvent interfaces in a poor solvent. Peptides within concentrated droplets can become indifferent to preferring intramolecular interactions to intermolecular interactions and consequently, individual chains can expand. In the ensuing conformational sampling, a presumed slow step, chains can converge on energetically favorable conformations that are high in β-content. Our proposal is similar to those of others for polyglutamine [35], Aβ [17] and the N-domain Sup35 [18] but needs to be tested using appropriate simulations.

3.10 Conclusion

The role of β-sheets in the early stages of protein aggregation, specifically amyloid formation, remains unclear. Interpretations of kinetic data have led to a specific model for the role of β-sheets in polyglutamine aggregation. According to this model, monomeric

polyglutamine, which is intrinsically disordered, goes through a rare conversion into an ordered, metastable, β-sheeted state that nucleates aggregation. It has also been proposed that the probability of forming the critical nucleus, a specific β-sheet conformation for the monomer increases with increasing chain length. Here, we test this model using molecular simulations. We quantified free energy profiles in terms of β-content for monomeric polyglutamine as a function of chain length. In accord with estimates from experimental data, the free energy penalties for forming β-rich states are in the 10-20 kcal / mol range. However, the length dependence of these free energy penalties does not mirror interpretations of kinetic data. Also, while homodimerization of disordered molecules is spontaneous, the imposition of conformational restraints on polyglutamine molecules does not enhance the spontaneity of intermolecular associations. Our data lead to the proposal that β-sheet formation is an attribute of peptide-rich phases such as high molecular weight aggregates rather than monomers or oligomers.

3.11 References

1. Vitalis, A., Lyle, N., and Pappu, R. V. (2009) Thermodynamics of beta sheet formation in polyglutamine, *Biophysical Journal 97*, 303-311.
2. Williams, A. J., and Paulson, H. L. (2008) Polyglutamine neurodegeneration: protein misfolding revisited, *Trends in Neurosciences 31*, 521-528.
3. Ross, C. A. (1995) When more is less: Pathogenesis of glutamine repeat neurodegenerative diseases, *Neuron 15*, 493-496.
4. Ross, C. A., and Poirier, M. A. (2004) Protein aggregation and neurodegenerative disease, *Nature Reviews Neuroscience*, S10-S17.

5. Chen, S. M., Berthelier, V., Hamilton, J. B., O'Nuallain, B., and Wetzel, R. (2002) Amyloid-like features of polyglutamine aggregates and their assembly kinetics, *Biochemistry 41*, 7391-7399.

6. Perutz, M. F., Finch, J. T., Berriman, J., and Lesk, A. (2002) Amyloid fibers are water-filled nanotubes, *Proceedings Of The National Academy Of Sciences Of The United States Of America 99*, 5591-5595.

7. Sikorski, P., and Atkins, E. (2005) New model for crystalline polyglutamine assemblies and their connection with amyloid fibrils, *Biomacromolecules 6*, 425-432.

8. Chen, S., Berthelier, V., Yang, W., and Wetzel, R. (2001) Polyglutamine aggregation behavior in vitro supports a recruitment mechanism of cytotoxicity, *Journal Of Molecular Biology 311*, 173-182.

9. Masino, L., Kelly, G., Leonard, K., Trottier, Y., and Pastore, A. (2002) Solution structure of polyglutamine tracts in GST-polyglutamine fusion proteins, *FEBS Letters 513*, 267-272.

10. Crick, S. L., Jayaraman, M., Frieden, C., Wetzel, R., and Pappu, R. V. (2006) Fluorescence correlation spectroscopy shows that monomeric polyglutamine molecules form collapsed structures in aqueous solutions, *Proceedings Of The National Academy Of Sciences Of The United States Of America 103*, 16764-16769.

11. Vitalis, A., Wang, X., and Pappu, R. V. (2007) Quantitative Characterization of Intrinsic Disorder in Polyglutamine: Insights from Analysis Based on Polymer Theories, *Biophysical Journal 93*, 1923-1937.

12. Chen, S. M., Ferrone, F. A., and Wetzel, R. (2002) Huntington's disease age-of-onset linked to polyglutamine aggregation nucleation, *Proceedings Of The National Academy Of Sciences Of The United States Of America 99*, 11884-11889.

13. Vitalis, A., Wang, X., and Pappu, R. V. (2008) Atomistic Simulations of the Effects of Polyglutamine Chain Length and Solvent Quality on Conformational Equilibria and Spontaneous Homodimerization, *Journal of Molecular Biology 384*, 279-297.

14. Vitalis, A., and Pappu, R. V. (2009) ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions, *Journal of Computational Chemistry 30*, 673-700.

15. Bhattacharyya, A. M., Thakur, A. K., and Wetzel, R. (2005) Polyglutamine aggregation nucleation: Thermodynamics of a highly unfavorable protein folding reaction, *Proceedings Of The National Academy Of Sciences Of The United States Of America 102*, 15400-15405.

16. Lomakin, A., Asherie, N., and Benedek, G. B. (2003) Liquid-solid transition in nuclei of protein crystals, *Proceedings of the National Academy of Sciences USA 100*, 10254-10257.

17. Bitan, G., Kirkitadze, M. D., Lomakin, A., Vollers, S. S., Benedek, G. B., and Teplow, D. B. (2003) Amyloid beta-protein (Abeta) assembly: Abeta40 and Abeta42 oligomerize through distinct pathways, *Proceedings of the National Academy of Sciences USA 100*, 330-335.

18. Krishnan, R., and Lindquist, S. L. (2005) Structural insights into a yeast prion illuminate nucleation and strain diversity, *Nature 435*, 765-772.

19. Engh, R. A., and Huber, R. (1991) Accurate Bond And Angle Parameters For X-Ray Protein-Structure Refinement, *Acta Crystallographica. Section A, Crystal Physics, Diffraction, Theoretical and General Crystallography 47*, 392-400.

20. Kaminski, G. A., Friesner, R. A., Tirado-Rives, J., and Jorgensen, W. L. (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides, *Journal Of Physical Chemistry B 105*, 6474-6487.

21. Still, W. C., Tempczyk, A., Hawley, R. C., and Hendrickson, T. (1990) Semianalytical Treatment Of Solvation For Molecular Mechanics And Dynamics, *Journal Of The American Chemical Society 112*, 6127-6129.

22. Gallicchio, E., and Levy, R. M. (2004) AGBNP: An analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling, *Journal Of Computational Chemistry 25*, 479-499.

23. Feig, M., Im, W., and Brooks, C. L. (2004) Implicit solvation based on generalized Born theory in different dielectric environments, *Journal Of Chemical Physics 120*, 903-911.

24. Lazaridis, T., and Karplus, M. (1999) Effective energy function for proteins in solution, *Proteins-Structure Function And Genetics 35*, 133-152.

25. Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., and Kollman, P. A. (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method, *Journal of Computational Chemistry 13*, 1011-1021.

26. Roux, B. (1995) Calculation of the potential of mean force using computer simulations, *Computer Physics Communications 91*, 275-282.

27. Hobohm, U., Scharf, M., Schneider, R., and Sander, S. (1992) Selection of a representative set of structures from the Brookhaven Protein Data Bank., *Protein Science 1*, 409-417.

28. Kabsch, W., and Sander, C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features., *Biopolymers 22*, 2577-2637.

29. Sugita, Y., Kitao, A., and Okamoto, Y. (2000) Multidimensional replica-exchange method for free-energy calculations, *Journal Of Chemical Physics 113*, 6042-6051.

30. Camacho, C. J., and Thirumalai, D. (1993) Minimum energy compact structures of random sequences of heteropolymers, *Physical Review Letters 71*, 2505-2508.

31. Humphrey, W., Dalke, A., and Schulten, K. (1996) VMD - Visual Molecular Dynamics, *Journal of Molecular Graphics 14*, 33-38.

32. Milchev, A., Paul, W., and Binder, K. (1993) Off-lattice Monte Carlo simulation of dilute and concentrated polymer solutions under theta conditions, *The Journal of Chemical Physics 99*, 4786-4798.

33. Marchut, A. J., and Hall, C. K. (2007) Effects of chain length on the aggregation of model polyglutamine peptides: Molecular dynamics simulations, *Proteins-Structure Function And Bioinformatics 66*, 96-109.

34. Barton, S., Jacak, R., Khare, S. D., Ding, F., and Dokholyan, N. V. (2007) The length dependence of the polyQ-mediated protein aggregation, *Journal of Biological Chemistry 282*, 25487-25492.

35. Lee, C. C., Walters, R. H., and Murphy, R. M. (2007) Reconsidering the mechanism of polyglutamine peptide aggregation, *Biochemistry 46*, 12810-12820.

36. Pappu, R. V., Wang, X., Vitalis, A., and Crick, S. L. (2008) A polymer physics perspective on driving forces and mechanisms for protein aggregation, *Archives of Biochemistry and Biophysics 469*, 132-141.

Chapter 4

Opposing effects of glutamine and asparagine dictate prion formation by intrinsically disordered proteins

This chapter is adapted from an article [1] published in Molecular Cell and the majority of the work presented here are from the article's co-authors in the Lindquist group. For completeness, the entirety of the results is included. Nicholas J. Lyle (NJL), the candidate, and Rohit V. Pappu (RVP) designed the simulations. NJL performed the simulations and analyzed the results. Randal Halfmann (RH), Simon Alberti (SA), and Susan Lindquist designed the experiments. RH, SA, Rajaraman Krishnan (RK), Charles W. O'Donnell (CWO), Oliver D. King (ODK), and Bonnie Berger (BB) performed the experiments. We also acknowledge Charles Glabe (U. C. Irvine) for providing the A11 polyclonal antibodies and thank the National Institutes of Health (GM025874 to SL and NS056114 to RVP) for funding.

4.1 Introduction

Protein regions with little to no regular structure – intrinsically disordered regions (IDRs) – are abundant in eukaryotic proteomes [2]. Such regions play important roles in gene regulation, signaling circuitries, and intracellular transport and are often centrally located in protein interaction networks [3, 4]. Many of these disordered regions recognize their macromolecular partners by undergoing disorder-to-order transitions. IDRs can also be susceptible to promiscuous interactions that pose a burden for cellular protein homeostasis [5, 6].

Some archetypal IDRs have low complexity amino acid sequences that are depleted of order-promoting residues [7]. Here, our interest is in a subset of low complexity sequences that are enriched in the polar uncharged residues glutamine (Q) and asparagine (N). Despite their general tendency toward promoting disorder at the monomer level [8-10], Q/N-rich sequences, can self-assemble into ordered hierarchical structures namely, amyloids [11-13], which are pseudocrystalline fibrillar assemblies stabilized by extensive intermolecular interactions between individual polypeptide monomers [14]. Thus, during amyloid formation, Q/N-rich proteins transition from one extreme of conformational space to the other. This transition appears to involve the initial formation of molten oligomers. The monomeric polypeptides form collapsed globules within these molten oligomers. Disorder-to-order transitions of individual polypeptides within these oligomers can be facilitated by intermolecular interactions and this can produce an amyloid-nucleating species [15-19].

Several protein misfolding diseases, including Huntington's disease and multiple spinocerebellar ataxias, are associated with the aggregation of polyQ sequences. Both the severity of the disease and the aggregation tendency of the protein are correlated with the length of the polyQ tract [20]. Q-rich and N-rich proteins can also undergo conformational switches to amyloid under non-pathological conditions, and in some cases these amyloids have important biological roles. Among such functional amyloids, extracellular adhesins of bacteria constitute an ancient and broadly distributed class of proteins that act as structural scaffolds for biofilm formation [21-23]. For these proteins, the initial switch from disorder to amyloid catalyzed by a dedicated nucleation machinery on the cell surface [24]. Another class of Q/N-rich proteins can serve as "protein only" elements of inheritance

when their IDRs switch, at a low frequency, to the amyloid state. The latter, comprising the majority of known prion proteins, are united only by the ability of their amyloid conformations to perpetuate through a self-templating protein folding reaction that heritably alters the functions of their associated globular domains [11, 25].

In the baker's yeast *Saccharomyces cerevisiae*, in which Q/N-rich prions have been characterized most extensively, the self-templating conformational conversions of prion proteins continues for multiple generations, and thereby causes robustly heritable phenotypes. Prion phenotypes can derive either from sequestration of the protein's globular domain, or from novel functions conferred by the prion conformation itself [26]. Unlike the well-known mammalian prion protein, PrP, which causes transmissible spongiform encephalopathies, yeast prions are not overtly toxic. In fact, the phenotypic diversity generated by yeast prion switching has been proposed to facilitate survival and adaptation in the rapidly changing natural environments of microbial cells [27, 28]. Whether yeast prions have evolved for this purpose remains to be established.

A recent genome-wide survey found that prion-forming proteins were more likely to be N-rich than Q-rich [11]. This observation was unexpected, as it challenged the common assumption that Ns and Qs are equivalent in facilitating prion formation [29, 30, 31]. However, it was subsequently suggested that the observed bias was better explained by a coupling between Q- vs. N-richness and local sequence context. This observation was made based on a set of experiments where structure-breaking proline residues were more abundant in the Q-rich sequences were part of the published study [9], leading to the proposal that it is prolyl residues within a Q-rich context rather than inherent differences between Q- and N-rich sequences that contribute to the biases observed by Alberti et al.

Here, we provide an in-depth analysis of the contributions of N and Q residues to prion formation. It is important to emphasize that these comparative assessments are performed in similar sequence contexts, i.e., for every sequence variant we compare the wild type sequence that has an admixture of Qs and Ns to variants where all the Qs / Ns are replaced by Qs and Ns, respectively. This allows us to focus on the differences between Qs and Ns in different sequence contexts and make clear adjudications about their intrinsic and context-dependent behaviors.

4.2 Qs and Ns have disparate effects on prion formation by Sup35

To compare the effects of Ns and Qs on prion formation, we generated two variants of the amyloidogenic prion domain (PrD) of the yeast prion protein Sup35. Normally, 15% of the residues in this PrD are Ns and 29% are Qs ($Sup35^{WT}$). In the two modified variants, either all Q residues were replaced with Ns ($Sup35^{N}$), or all N residues were replaced with Qs ($Sup35^{Q}$) (Figure 4.1A). The sequences were otherwise identical.

We first analyzed the propensity of these proteins to form alternative self-propagating conformations *in vivo*, using a phenotypic reporter for prion formation that depends on changes in the activity of Sup35 [11]. Sup35 is a translation termination factor. When the prion domain switches to the assembled amyloid state, it sequesters Sup35 from ribosomes, causing them to read through stop codons at an increased frequency. In cells that have an *ADE1* gene with a premature stop codon, read through changes colony color from red to white and allows growth without adenine. Each Sup35 variant ($Sup35^{WT}$, $Sup35^{N}$, and $Sup35^{Q}$) was constitutively expressed in strains lacking endogenous Sup35. They accumulated to similar levels (Figure 4.2B) and yielded a

comparable red colony color. That is, all possessed normal Sup35 activity and could stably maintain a soluble non-prion state (Figure 4.2C).

Normally, the spontaneous rate of prion formation by Sup35 is quite low, ~ one in 106 cells per generation [32]. To allow a better quantitative comparison between variants, we increased the likelihood of prion conversion by over-expressing each PrD variant [11, 33] as an EYFP fusion from an inducible promoter (*GAL1*, Figure 4.2A). This caused the expected increase in white Ade+ colonies for $Sup35^{WT}$. For $Sup35^{N}$, white Ade+ colonies were slightly more frequent. For $Sup35^{Q}$ they were essentially absent (Figure 4.1B).

To confirm that white Ade+ colonies were due to prion formation, we first tested their dependence on Hsp104, a AAA+ ATPase whose amyloid-fragmenting activity is critical for prion propagation [34]. We passed presumptive prion colonies on media containing a low concentration of guanidine hydrochloride (GdnHCl), which selectively inhibits the ATPase activity of Hsp104 [35]. This restored $Sup35^{WT}$ and $Sup35^{N}$ cells to their original red phenotypes (Figure 4.1C). Genetic ablation of *HSP104* had the same effect (Figure 4.1C). Second, we asked if nonsense suppressing $Sup35^{N}$ cells contained the SDS-resistant amyloids that are the prion template. Semi-denaturing detergent-agarose gel electrophoresis (SDD-AGE) confirmed that they did. Indeed, $Sup35^{N}$ cells gave rise to a range of phenotypically distinct prion states (white and pink color variants) [33] typical of the wild-type protein, and these were associated with the expected differences in amyloid size [36] (Figure 4.1D, 2E).

Another regulator of the appearance of yeast prions is the prion-inducing factor known as [*RNQ*+]. This factor is itself a prion conformer of the Rnq1 protein, hence the brackets (cytoplasmic inheritance) and capital letters (genetic dominance) in its

nomenclature. The [*RNQ*+] prion strongly enhances the ability of other proteins, like Sup35, to convert *de novo* to their own prion states [37]. But once these other proteins have formed self-propagating prion elements, they no longer require [*RNQ*+]. We eliminated [*RNQ*+] by deletion of the *RNQ1* gene. In $Sup35^{WT}$ and $Sup35^{N}$ cells that had already acquired nonsense-suppression phenotypes, this deletion had no effect (Figure 4.1C). As expected, this deletion blocked the *de novo* induction of nonsense suppression phenotypes in $Sup35^{WT}$ cells. However, while it reduced the appearance of prions in $Sup35^{N}$ cells, it did not eliminate it (Figure 4.2D). The unique ability of $Sup35^{N}$ to form prions independently of [RNQ+] suggests that it has an increased basal tendency to form amyloid.

The rare white colonies that appeared with $Sup35^{Q}$ proved not to be due to the formation of prions. Subsequent experiments revealed that rare colonies with all the hallmarks of prions could form when the protein was expressed at extremely high levels (Figure 4.2G). These states, however, were highly unstable when expression was reduced (not shown). Thus, shifting the Sup35 prion sequence to a more N-rich form increased its prion-forming capacity to such an extent that it bypassed the normal requirement for prion-inducing factor. Shifting the sequence to a more Q-rich form virtually eliminated its ability to form prions.

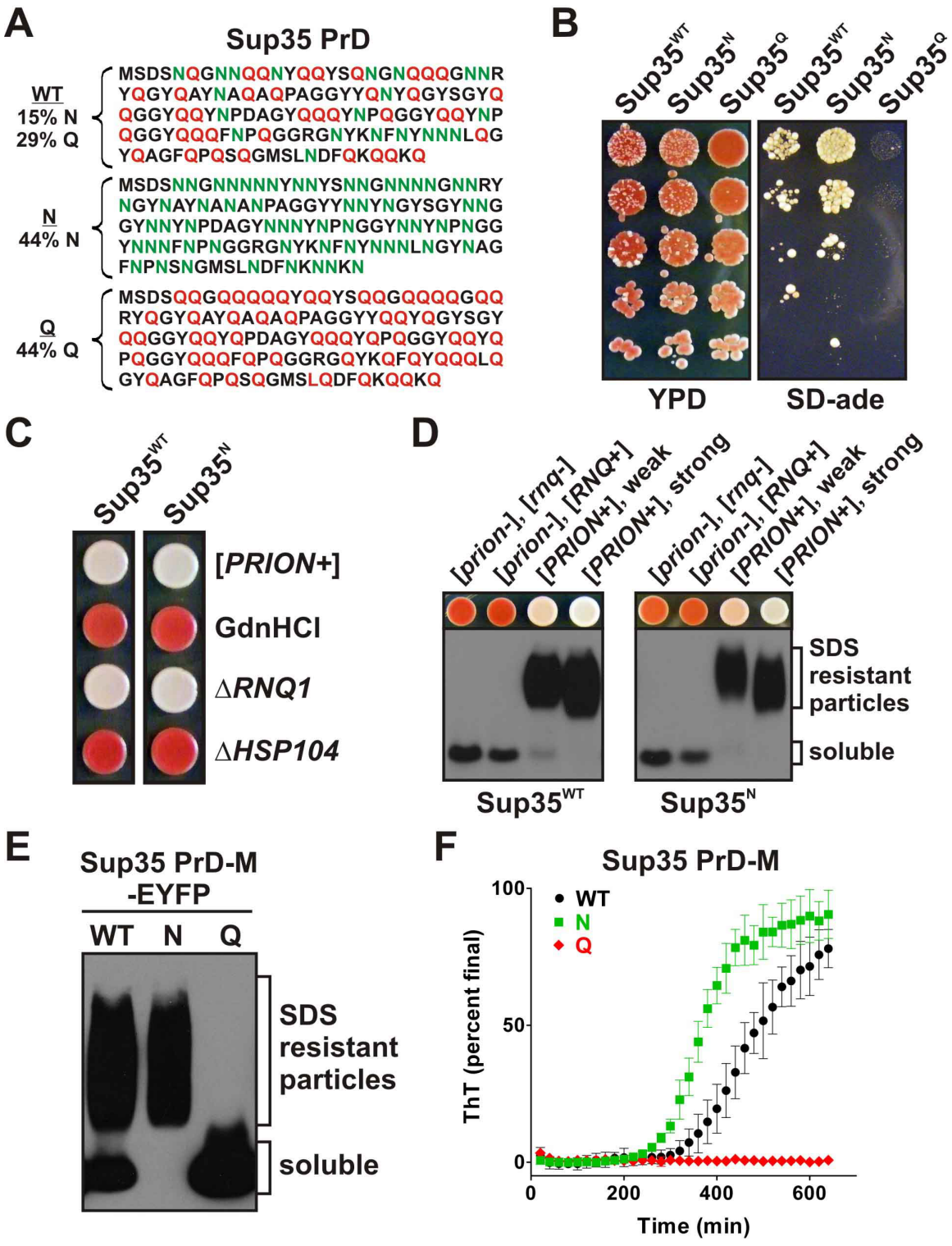

A
Sup35 PrD
WT
15% N
29% Q
MSDSNQGNNQQNYQQYSQNGNQQQGNNR
YQGYQAYNAQAQPAGGYYQNYQGYSGYQ
QGGYQQYNPDAGYQQQYNPQGGYQQYNP
QGGYQQQFNPQGGRGNYKNFNYNNNLQG
YQAGFQPQSQGMSLNDFQKQQKQ
N
44% N
MSDSNNGNNNNNYNNYSNNGNNNNGNNRY
NGYNAYNANANPAGGYYNNYNGYSGYNNG
GYNNYNPDAGYNNNYNPNGGYNNYNPNGG
YNNNFNPNGGRGNYKNFNYNNNLNGYNAG
FNPNSNGMSLNDFNKNNKN
Q
44% Q
MSDSQQGQQQQQYQQYSQQGQQQQGQQ
RYQGYQAYQAQAQPAGGYYQQYQGYSGY
QQGGYQQYQPDAGYQQQYQPQGGYQQYQ
PQGGYQQQFQPQGGRGQYKQFQYQQQLQ
GYQAGFQPQSQGMSLQDFQKQQKQ
B
Sup35WT
Sup35N
Sup35Q
Sup35WT
Sup35N
Sup35Q
YPD
SD-ade
C
Sup35WT
Sup35N
[PRION+]
GdnHCl
ΔRNQ1
ΔHSP104
D
[prion-], [rnq-]
[prion-], [RNQ+]
[PRION+], weak
[PRION+], strong
[prion-], [rnq-]
[prion-], [RNQ+]
[PRION+], weak
[PRION+], strong
SDS resistant particles
soluble
Sup35WT
Sup35N
E
Sup35 PrD-M -EYFP
WT
N
Q
SDS resistant particles
soluble
F
Sup35 PrD-M
WT
N
Q
ThT (percent final)
100
50
0
0
200
400
600
Time (min)

Figure 4.1: Prion formation by Sup35 is promoted by Ns, inhibited by Qs. (A) WT sequence of the Sup35 PrD (top), and Q and N replacement variants. (B) Yeast strains expressing Sup35 variants spotted as 5-fold serial dilutions onto YPD (nonselective) or SD-ade (prion-selective) plates. Prion states were induced by the over-expression of PrD-M-EYFP fusions for 24 hours prior to plating. (C) The N-substituted variant of Sup35 can form a prion state that is equivalent to that of WT. White Ade+ $Sup35^{N}$ cells were isolated and passaged on plates containing 5 mM GdnHCl or transformed with gene-specific knock-out cassettes to delete *RNQ1* ("ΔRNQ1") or *HSP104* ("ΔHSP104"). All presumptive prion strains were curable and lost the prion state upon deletion of *HSP104*. A representative [*PRION*+] strain of $Sup35^{N}$ (right) is compared to a strong [*PRION*+] strain of $Sup35^{WT}$ (left). (D) $Sup35^{N}$ can form different conformational variants that are equivalent to those of $Sup35^{WT}$. Colonies with weak and strong Ade+ phenotypes were isolated (Figure 4.2E). SDS-resistant aggregates were detected by SDD-AGE and immunoblotting with a Sup35C-specific antibody. (E) Variant Sup35 PrD-M-EYFP fusions were expressed for 24 hours in [*RNQ*+] cells prior to SDD-AGE analysis. PrD-M-EYFP was detected with a GFP-specific antibody. (F) Sup35 PrD-M-His7 variants were purified under denaturing conditions and then diluted to 5 μM in assembly buffer. Reactions were agitated for 10 sec every 2 min in the presence of non-binding plastic beads. Amyloid formation was monitored by ThT fluorescence. Data were normalized by the final values achieved for each variant after extended incubations. Data represent means +/- SEM.

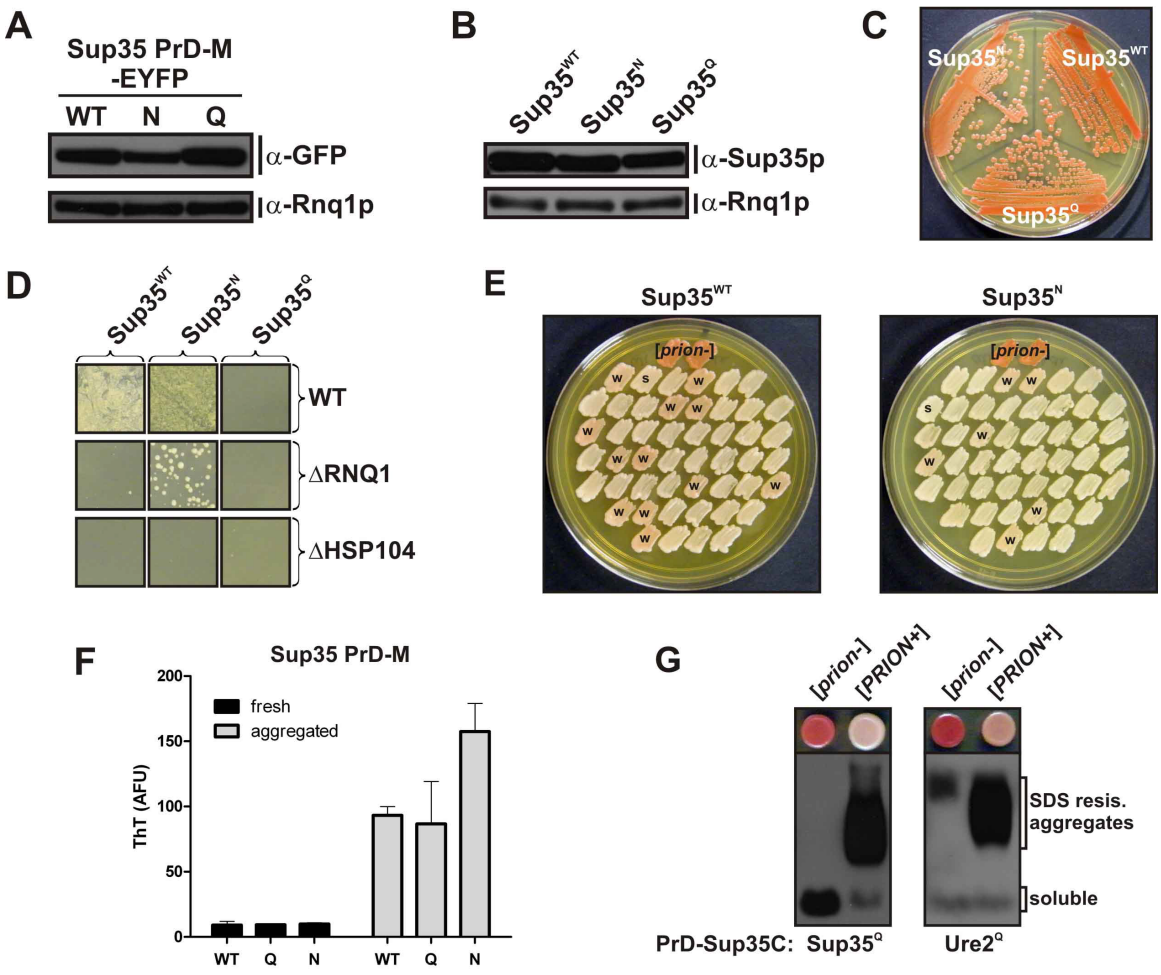


Figure 4.2: Transformants possessed normal Sup35 activity and could stably maintain a soluble non-prion state. (A) Constructs expressing the indicated proteins from a galactose-regulatable promoter were introduced into yeast cells. The resulting transformants were grown for 24 hours under inducing conditions (in the presence of galactose) and cell lysates were prepared and subjected to SDS-PAGE. Separated proteins were analyzed by immunoblotting with a GFP-specific antibody. Detection of Rnq1 with a Rnq1-specific antibody was used to demonstrate equal loading. (B) Cell lysates of yeast cells expressing $Sup35^{WT}$, $Sup35^{N}$ and $Sup35^{Q}$ were analyzed by immunoblotting with a Sup35-specific antibody. Equal loading was demonstrated by

immunodetection of Rnq1 with a Rnq1-specific antibody. (C) Yeast cells expressing $Sup35^{WT}$, $Sup35^{N}$ and $Sup35^{Q}$ display comparable colony colors when in the [*prion*-] states, indicating that the variant fusion proteins are soluble and have no defects in translation termination. (D) Yeast strains containing variant Sup35 differ in the prion induction frequency and their dependence on [*RNQ*+]. $Sup35^{WT}$, $Sup35^{N}$ and $Sup35^{Q}$ yeast were grown to mid-log phase, normalized by OD_{600} and plated onto SD-ade plates. Prion states were induced prior to plating by overexpression of PrD-M-EYFP fusions for 24 hours. Yeast cells deleted for *RNQ1* (ΔRNQ1) and *HSP104* (ΔHSP104) were compared to wild-type cells (WT). (E) $Sup35^{N}$ can form prion strains that are reminiscent of strong and weak [*PSI*+]. The prion state was induced by expression of $Sup35^{N}$ PrD-M-EYFP for 24 hours and the cells were subsequently plated onto adenine-deficient medium to select for prions. Ade+ colonies were isolated and transferred to YPD plates. Weak and strong prion variants that resulted from *de novo* induction of [*PSI*+] are shown for comparison (left). (F) Purified denatured Sup35 PrD-M-His7 variants diluted to 5 μM in assembly buffer were incubated with-end-over rotation for 5 days and then examined for ThT-fluorescence (error bars = SEM). (G) ADH1 promoter-driven expression constructs for Sup35PrD-M-Sup35C or Ure2PrD-Sup35C were introduced into yeast cells to replace the endogenous *SUP35* gene. The resulting strains were transformed with plasmids expressing $Sup35^{Q}$ PrD-M-EYFP or $Ure2^{Q}$ PrD-EYFP from the strong GPD promoter. To select for prion states the cells were then plated onto adenine-deficient medium and Ade+ colonies were transferred onto YPD plates. Yeast strains showing a white colony color on YPD were lysed and the resulting lysates were analyzed by SDD-AGE and immunoblotting with a Sup35-specific antibody.

4.3 Q and N have disparate effects on amyloid formation by Sup35

Next we asked if the Q and N protein variants had intrinsically different propensities to form amyloid *de novo*. We induced Sup35 PrD-EYFP variants for 24 hours in cells that did not carry prion forms of the proteins. Despite similar expression levels (Figure 4.2A), these variants showed very different behaviors (Figure 4.1E). The WT protein partitioned between SDS-soluble and amyloid states. All detectable $Sup35^{N}$ coalesced into SDS-resistant polymers. All of the $Sup35^{Q}$ remained SDS-soluble.

We next asked if these differences depend upon the cellular environment, or if they reflect inherently different properties of each variant? We purified the variants from bacteria under conditions that were fully denaturing. The proteins were then diluted into a physiological assembly buffer containing Thioflavin-T (ThT), a dye that fluoresces upon binding amyloid [(38)]. $Sup35^{WT}$ and $Sup35^{N}$ formed ThT-binding species after a short lag phase, as is characteristic for prion proteins, with $Sup35^{N}$ achieving this state more rapidly than $Sup35^{WT}$. $Sup35^{Q}$ was incapable of forming amyloids in the same time frame (Figure 4.1F, 2F).

4.4 Ns and Qs influence other proteins in qualitatively similar ways

We also probed the effects of sequence context on the distinction between Qs and Ns for prion formation. It was previously noted that Q-rich sequences tend also to be enriched for prolines [(9)], and this might explain the apparent bias against Qs in diverse

prion-forming sequences [11]. We find, however, that even when proline-density is accounted for in these sequences, Ns are still more prionogenic than Qs (Figure 4.4A-B).

To further determine if the effects of Ns and Qs were generalizable, we created N→Q variants of two highly N-rich PrDs, one from Ure2 and the other from Lsm4 ($Ure2^{Q}$ and $Lsm4^{Q}$, Figure 4.3A), and subjected them to the same tests used for Sup35. Using Sup35C-fusions, $Ure2^{WT}$ and $Lsm4^{WT}$ drove prion formation at high frequencies. The corresponding Q-rich versions did not (Figure 4.3B, 4C-E). The Q-rich PrDs were also severely impaired for amyloid formation, both *in vivo* when over-expressed as EYFP fusions (Figure 4.3C) and *in vitro* following their purification and dilution into physiological buffer (Figure 4.3D). While the qualitative aspects of the comparative studies between Q- / N- variants of Ure2 and Lsm4 are similar in flavor to the results for Sup35, there are quantitative differences that one can assign to differences in sequence contexts. For example, comparisons between Figures 1F and 3D show differences in lag times and the time required to achieve 50% of the maximal ThT fluorescence (t_{50}). Specifically, the lag times and t_{50} values are longer for Ure2 and Lsm4 when compared to $Sup35^{N}$ and $Sup35^{WT}$, respectively. Furthermore, $Sup35^{WT}$ has more Qs than Ns, the lag times and yet the t_{50} values for this sequence are lower when compared to both Ure2 and Lsm4, both of which have more Ns than Qs. Cursory inspection reveals key differences in the sequence contexts for Ns / Qs in Ure2 and Lsm4 versus Sup35. The latter sequence has four putative turn-forming motifs *viz*., NPDA and NPQG that aren't altered in either $Sup35^{N}$ or $Sup35^{Q}$. It appears that the presence of these motifs might explain the comparable amyloid formation propensities of $Sup35^{N}$ and $Sup35^{WT}$. However, it is also clear that despite the presence of these turn forming motifs in $Sup35^{Q}$, the substitution of

Ns by Qs significantly diminishes amyloid formation. Apparently, within the sequence context of Sup35^{WT} there is a need for a balance between the ability to nucleate turns and the ability to propagate this into strands and this might be compromised by Q- as opposed to N-rich / hybrid Q- / N-rich tracts.

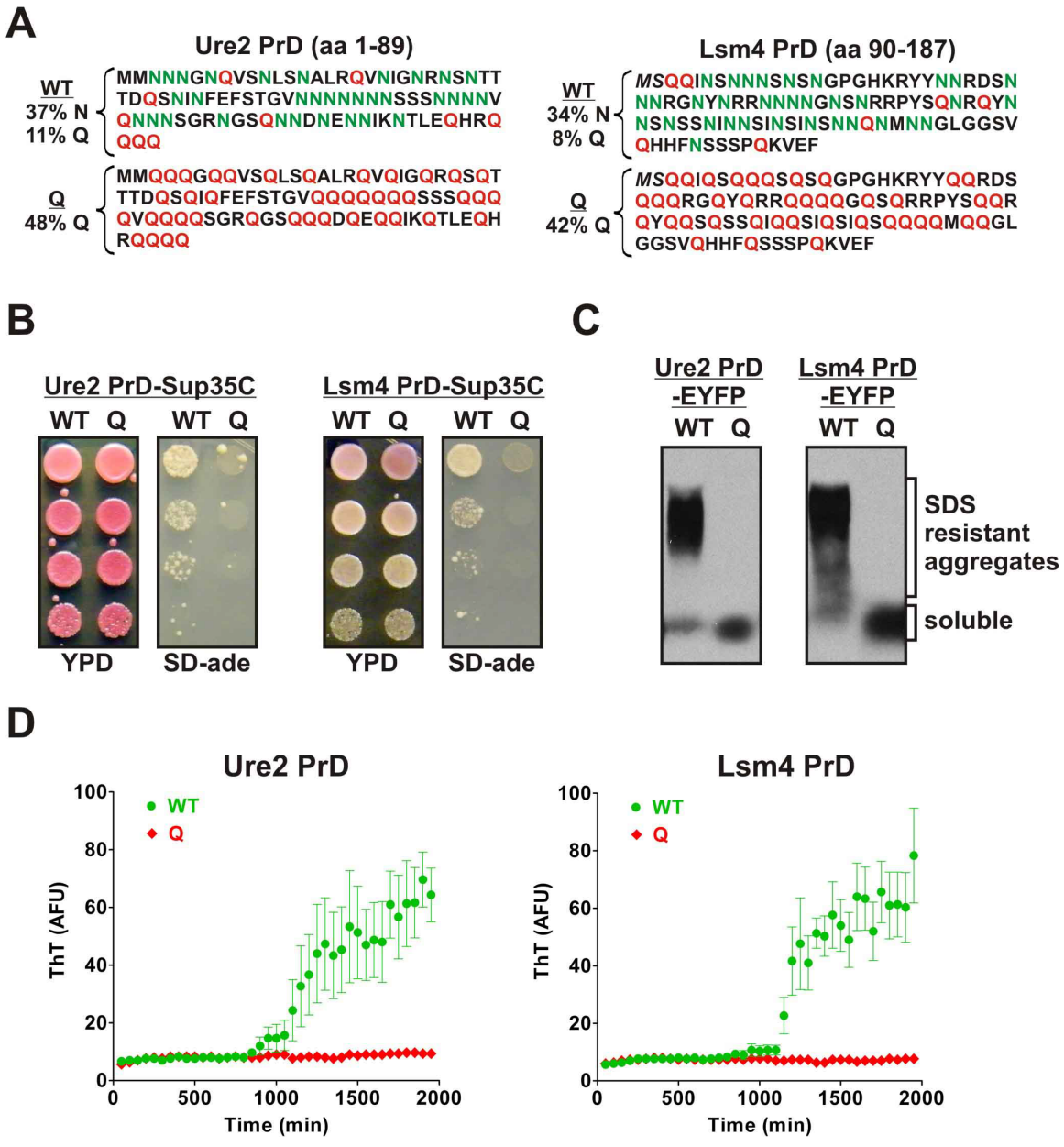


Figure 4.3: Replacing Ns with Qs eliminates prion-formation by N-rich PrDs. (A) The sequences of the Ure2 and Lsm4 PrDs (top), along with the Q variants. (B) Yeast strains containing variant Ure2 and Lsm4 PrDs fused to Sup35C were spotted to YPD and SD-ade plates as in Figure 4.1B. Prion states were induced by over-expression of PrD-EYFP fusions for 24 hours prior to plating. Representative Ade+ colonies for $Ure2^{WT}$ and $Lsm4^{WT}$ (but not the few Ade+ colonies observed for $Ure2^{Q}$) showed SDS-resistant aggregates by SDD-AGE and were eliminated by growth on GdnHCl (not shown). (C) Variant Ure2 and Lsm4 PrD-EYFP fusions were expressed for 24 hours in [*RNQ*+] cells prior to SDD-AGE analysis as in Figure 4.1E. (D) Purified denatured variants of Ure2 and Lsm4 PrD-His7 were diluted to 20 μM or 5uM, respectively, in assembly buffer. Reactions were agitated for 10 sec every 2 min in the absence of beads. Amyloid formation was monitored by ThT fluorescence. Data represent means +/- SEM.

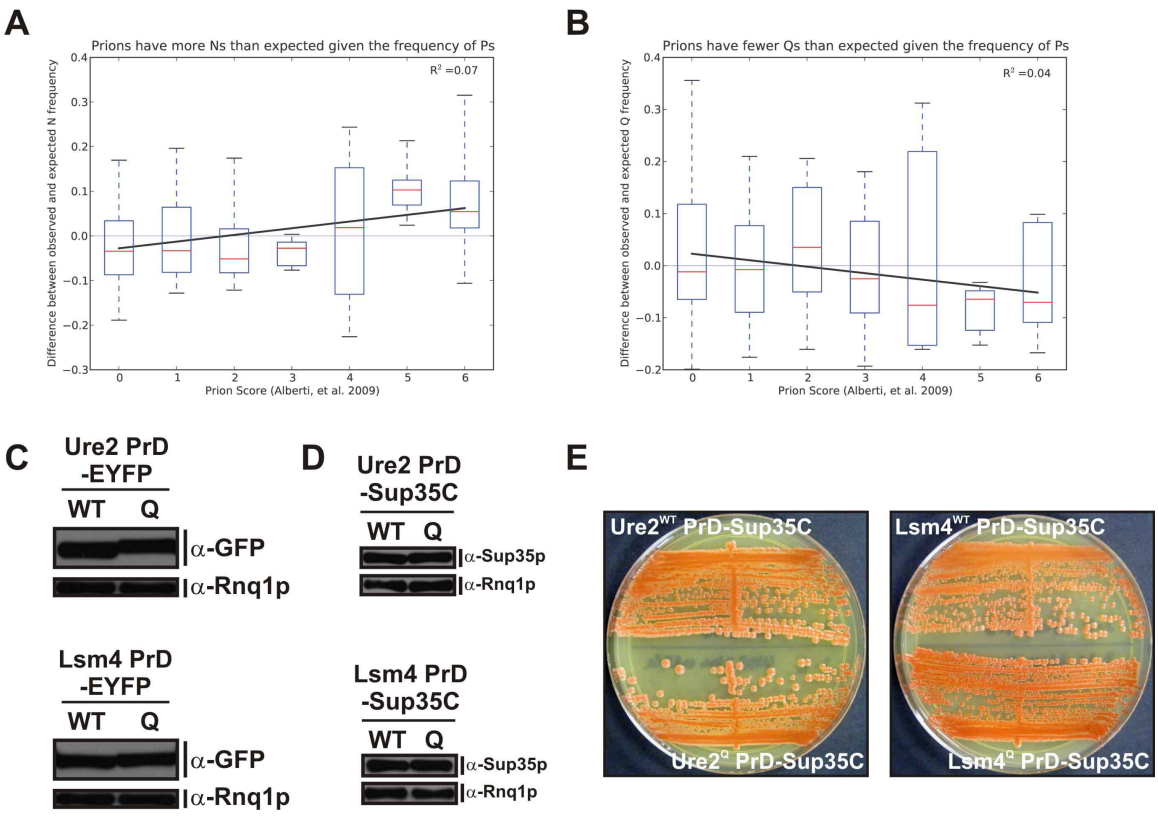


Figure 4.4: Replacing Ns with Qs eliminates prion-formation by N-rich PrDs. (A) Prions exhibit more Ns and fewer Qs than expected given the sequence frequency of proline (P). The 92 Q/N-rich sequences were clustered by their "prion score" as defined [(11)]. On the y-axis, the difference between the frequency of N and that which is expected given the sequence frequency of P, determined by linear regression of N and P frequencies in the total set of 92 sequences. (B) Similarly, the y-axis depicts the difference between the frequency of Q and that which is expected given the sequence frequency of P. Boxes include upper and lower quartiles; median in red; whiskers cover full range of values. (C) As in Figure 4.2A, for Ure2 and Lsm4 PrD-EYFP variants. (D) As in Figure 4.2B, for Ure2 and Lsm4 PrD-Sup35C variants. (E) As in Figure 4.2C, for Ure2 and Lsm4 PrD-Sup35C variants.

To determine if N-richness can drive prion formation in a protein that does not normally form them, we generated a Q→N variant of a fragment of the Gal11 transcription factor ($Gal11^N$, Figure 4.5A). As reported previously [11], the WT sequence lacks prion activity (Figure 4.5B, 6A-C). $Gal11^N$ readily produced Ade+ colonies, and these can be attributed to prion formation. The Ade+ phenotype was reversed by Hsp104 inactivation (Figure 4.5C) and, as expected for a prion, did not require the continued presence of [*RNQ*+] (Figure 4.5C). Moreover, SDD-AGE revealed SDS-insoluble $Gal11^N$-Sup35C amyloids in Ade+ cells (Figure 4.5D). The proteins were poorly expressed in *E. coli* for purification and *in vitro* analysis. However, the prion propensities of the Gal11 PrD variants corresponded to their amyloid propensities upon *de novo* expression in yeast cells (Figure 4.5E).

Many proteins associated with amyloid diseases contain long glutamine tracts (polyQ) with a propensity to aggregate in both the human brain [39] and when heterologously expressed in yeast [40-42]. To explore the distinction between Qs and Ns in such a polypeptide, we compared a disease-associated version of Htt exon 1 (Htt^{Q47}), with a Q→N variant of the same protein (Htt^{N47}, Figure 4.5F). When fused to EYFP and expressed for 24 hours, both variants formed SDS-resistant aggregates that were strongly promoted by the presence of [*RNQ*+] (Figure 4.5G). However, regardless of [*RNQ*+] status, Htt^{N47} partitioned much more completely to the SDS-resistant fraction than did Htt^{Q47}.

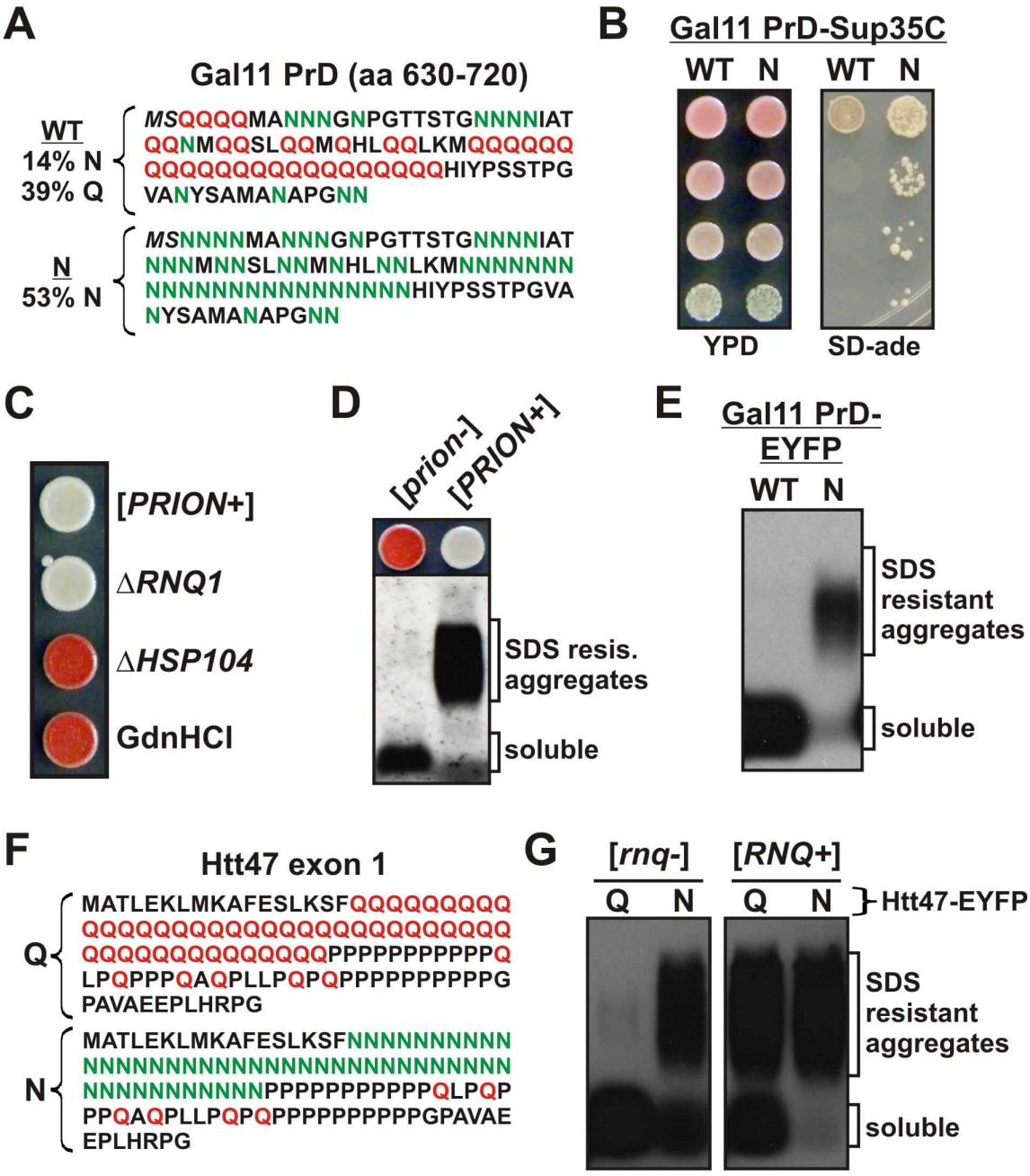


Figure 4.5: Replacing Qs with Ns increases amyloid and prion formation by Q-rich proteins. (A) WT and N variants of the putative PrD of Gal11, residues 630-720. (B) Yeast strains containing variants of the Gal11 PrD fused to Sup35C were spotted to YPD and SD-ade plates as in Figure 4.1B. Prion states were induced by over-expression of PrD-EYFP fusions for 24 hours prior to plating. (C-D) Gal11^{N} PrD-Sup35C-expressing

cells can convert to a prion state. Representative Ade+ cells were isolated and analyzed as in Figure 4.1C-D. (E) Variant PrD-M-EYFP fusions were expressed for 24 hours in [*RNQ*+] cells, followed by SDD-AGE analysis as in Figure 4.1D. (F) The sequence of Huntingtin exon 1 with a homopolymeric expansion of 47 Qs (top), and the N variant (bottom). (G) HttQ47 and HttN47 fused to EYFP were expressed for 24 hours in [*rnq*-] or [*RNQ*+] cells, followed by SDD-AGE analysis as in Figure 4.1E.

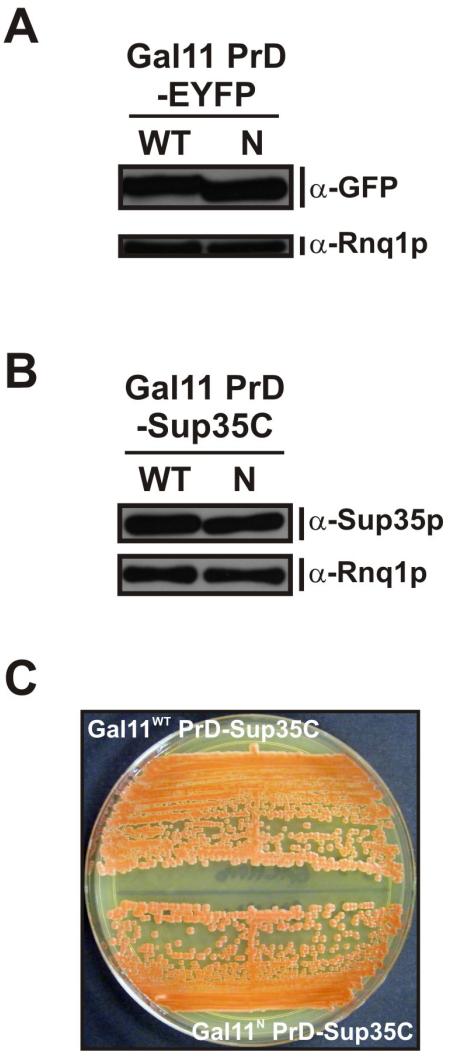


Figure 4.6, related to Figure 4.5. (A) As in Figure 4.2A, for Gal11 PrD-EYFP variants. (D) As in Figure 4.2B, for Gal11 PrD-Sup35C variants. (E) As in Figure 4.2C, for Gal11 PrD-Sup35C variants.

4.5 N-richness reduces proteotoxicity of Q/N-rich proteins

Over-expressed fluorescently-tagged prion proteins typically form bright puncti or ribbon-like foci indicative of bundled amyloid filaments [(11, 43, 44)]. Non prion-forming proteins remain diffuse or coalesce weakly into amorphous foci [(11)]. After 48 hours of expression from a galactose-inducible promoter, all of the N-rich, prion-proficient proteins studied here formed foci with sharp boundaries and often had an elongated filament-like morphology (Figure 4.7A, top). The Q-rich proteins also formed foci but these were less crisp and surrounded by diffuse fluorescence (Figure 4.7A, bottom). Note that the distinction we are making is not in the number of foci formed, but rather the extent to which they deplete soluble protein. By SDD-AGE, the N-rich proteins had acquired an SDS-resistant aggregated state whereas the Q-rich proteins were largely SDS-sensitive (Figure 4.8A-D).

Disordered proteins tend to be toxic when over-expressed. Our computational analyses of previously published data [(5)] indicate that this tendency is further increased by Q-richness (Figure 4.8E). To examine how aggregation tendencies of disordered proteins relate to toxicity, we transformed our galactose-inducible constructs into yeast cells carrying a chromosomal deletion of the PrD of Sup35. Cells were then spotted onto media that either induced or repressed expression. The Ure2, Lsm4, and Gal11 variants were not toxic (not shown). Of the Sup35 variants, $Sup35^{Q}$ and $Sup35^{WT}$ were mildly

toxic, whereas Sup35^{N} was benign (Figure 4.7B, left). Similarly, Htt^{Q47} was toxic relative to Htt^{N47} (Figure 4.7C, left).

To determine whether toxicity is enhanced or reduced by amyloid formation, we examined isogenic strains containing the amyloid-promoting factor [*RNQ*+]. This background decreased the toxicity of Sup35^{WT} (Figure 4.7B) as well as both variants of Htt (Figure 4.7C), consistent with the hypothesis that toxic activities of non-amyloid conformers are suppressed by amyloid formation. Notably, the toxicity of Sup35^{Q} was not reduced by [*RNQ*+] (Figure 4.7B), presumably because this variant lacks amyloidogenicity even in the presence of [*RNQ*+] (Figure 4.2D and data not shown).

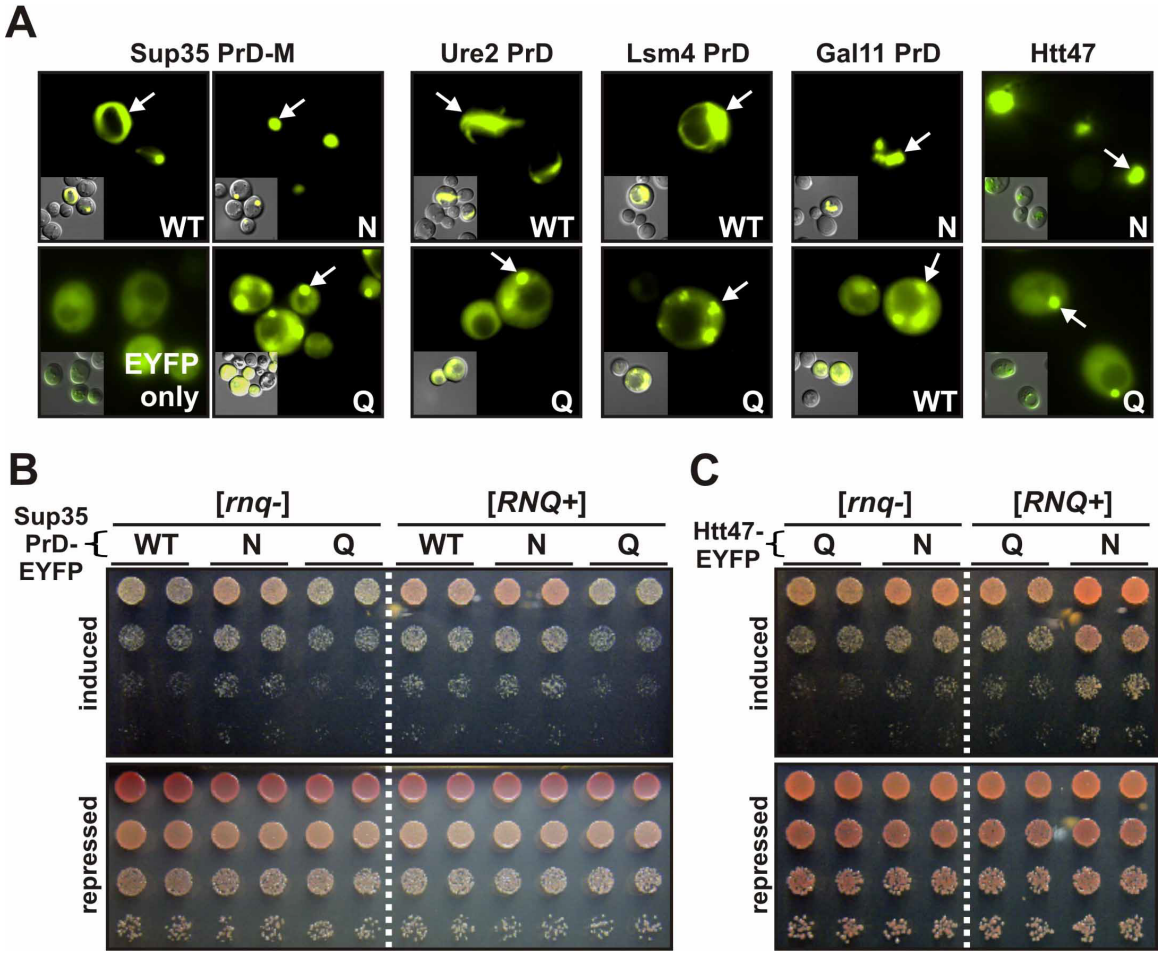


Figure 4.7: N-richness reduces proteotoxicity of Q/N-rich proteins. (A) Single-copy plasmids coding for PrD-EYFP fusions were introduced into [*RNQ*+] cells. Expression was induced by addition of galactose for 48 hours and protein localization was determined by fluorescence microscopy. (B) Isogenic [*rnq*-] or [*RNQ*+] yeast bearing the indicated Sup35 PrD-EYFP variants were spotted as 5-fold serial dilutions to plates that either induced (galactose) or repressed (glucose). Growth on glucose established that equal cell densities were plated for each variant. Differences in growth on galactose indicate toxicity resulting from expression of the indicated protein. Duplicate transformants are shown. White dotted lines are provided only for clarity; comparisons are made between cells growing on the same plate. (C) As in (B), but with HttQ47- and HttN47-EYFP.

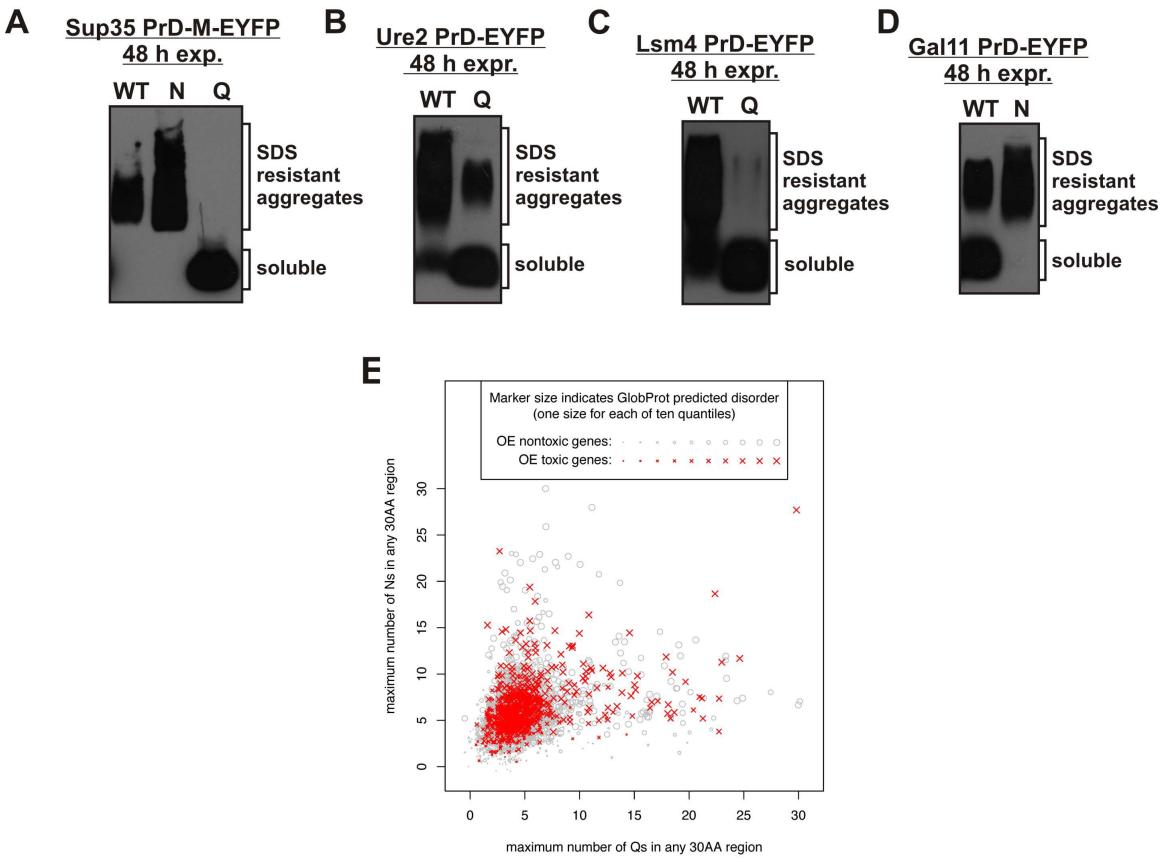


Figure 4.8, related to Figure 4.7. (A-D) Yeast cells expressing the indicated PrD-EYFP proteins were grown for 48 hours under inducing conditions and subjected to SDD-AGE, as in Figure 4.1E. (E) Correlation of Q- and N-richness with dosage sensitivity. Shown is a scatter plot depicting the relation between Q-richness (horizontal axis), N-richness (vertical axis), dosage sensitivity (red x for OE toxic proteins, grey o for OE nontoxic proteins) and disorder (indicated by marker size). Marker sizes increase with disorder, binned into 10 quantiles, so that e.g. the smallest size is for the 10% of proteins with the fewest disordered residues and the largest size is for the 10% of proteins with the most disordered residues. Coordinates of points were jittered by up to 0.5 in each direction to reduce overlap. (F) Growth curves of isogenic [*rnq*-] or [*RNQ*+] yeast bearing the

indicated Htt-EYFP variants in liquid inducing media (containing galactose). Growth was measured by absorbance at 600nm. Shown are means +/- SEM from three transformants each.

4.6 Q-rich proteins preferentially form toxic non-amyloid conformers

To investigate the inherent tendency of the Sup35 variants to partition between amyloid and non-amyloid states, we incubated purified proteins in assembly buffer for 24 hours, with end-over-end agitation. After centrifugation to collect aggregates, SDS was added, followed by a second centrifugation step. $Sup35^{WT}$ and $Sup35^{N}$ converted almost entirely to SDS-resistant amyloids (Figure 4.9A). A large fraction of $Sup35^{Q}$ remained soluble. The precipitating fraction was mostly SDS-soluble.

To determine if the different conformers formed by the Sup35 variants are inherently toxic, we applied purified protein preparations to human neuroblastoma cells in culture. None were toxic when freshly diluted from denaturant (as shown for $Sup35^{Q}$; Fig. 9B). When the proteins were allowed to aggregate for 24 hours, $Sup35^{Q}$ became severely toxic, causing membrane permeabilization (quantified by adenylate kinase release; Figure 4.9B) and cell detachment (Figure 4.9C). $Sup35^{N}$ and $Sup35^{WT}$ became only mildly toxic. The extreme distinctions between Sup35 variants in this assay prompted us to examine the Ure2 PrD variants as well. After 24 hours of aggregation, $Ure2^{WT}$ was only mildly toxic whereas $Ure2^{Q}$ was severely toxic (Figure 4.10).

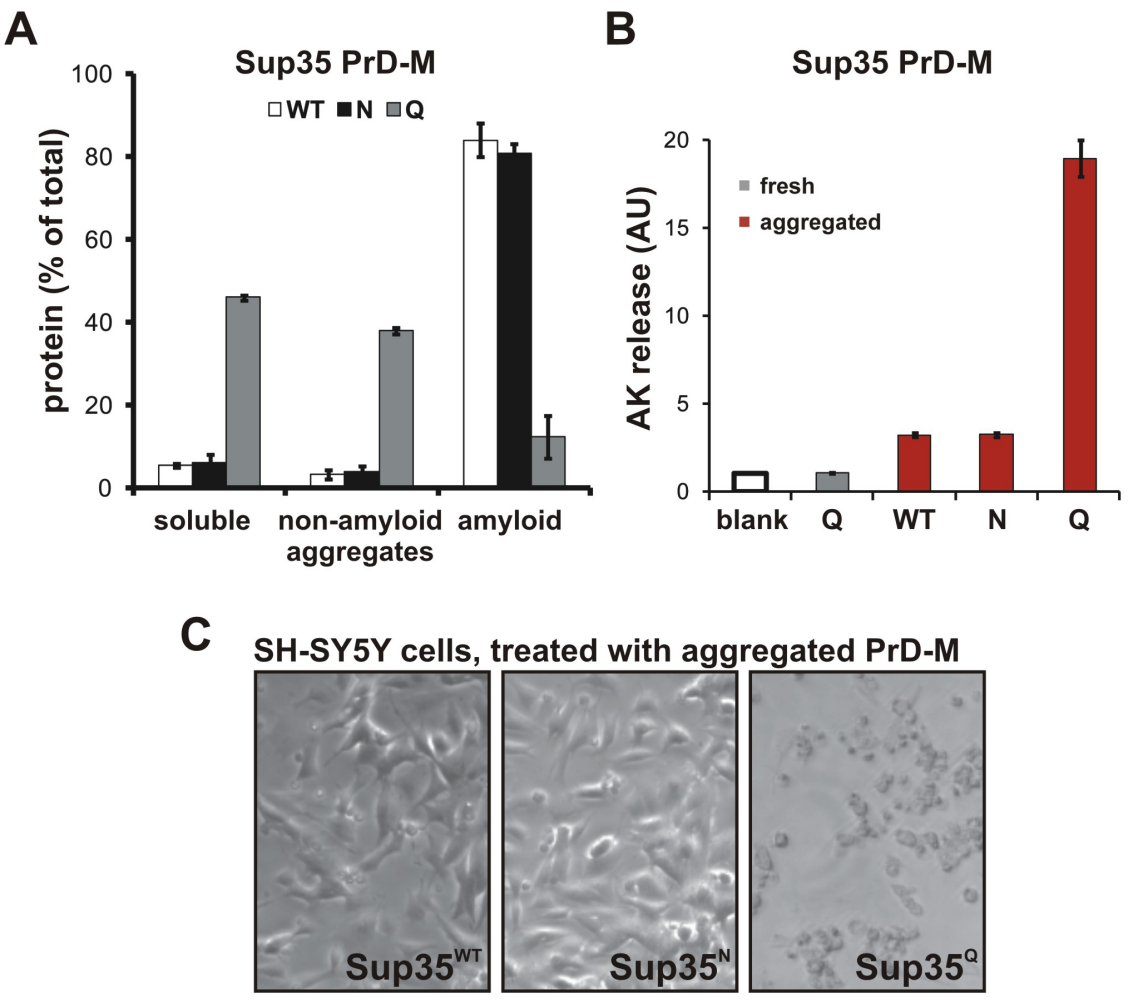


Figure 4.9: Q-rich proteins preferentially form non-amyloid conformers. (A) Quantitation of soluble, amyloid, and non-amyloid aggregated protein in assemblies of Sup35 PrD-M-His7 variants. Freshly diluted 5 μM solutions were induced to assemble with end-over-end agitation for 24 hours. Soluble and aggregated fractions were partitioned by centrifugation at 39,000 rcf for 30 min. The aggregate fraction was further resuspended in 1 % SDS and allowed to incubate at 25°C for 30 min, followed by a second centrifugation step. Protein concentrations are shown (+/- SEM) for the original supernatant ("soluble"), post-SDS supernatant ("non-amyloid aggregation") and post-SDS pellet ("amyloid aggregation"). (B-C) Toxicity of variant Sup35 PrD-M-His7

assemblies to human neuroblastoma cells. SH-SY5Y cells incubated for 15 hours with 2.5 μM of either freshly diluted or pre-aggregated protein, as indicated, were visually inspected for cell detachment (B) or assayed for membrane disruption by adenylate kinase release (C).

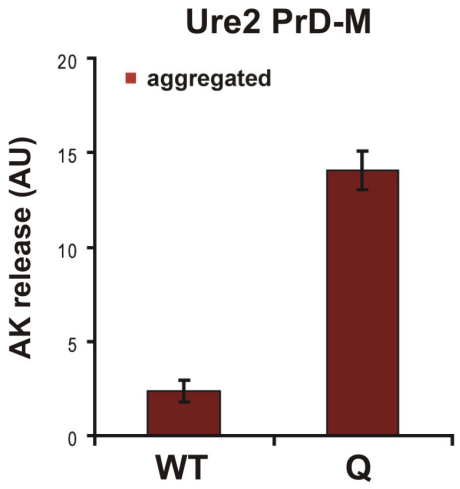


Figure 4.10: related to Figure 4.9. Toxicity of Ure2 PrD-M-His7 aggregates to SH-SY5Y cells, as for Sup35 in Figure 4.7D.

4.7 Q-rich proteins have a defect in amyloid conversion

What aspects of amyloid formation govern the distinct behaviors of these proteins? According to one set of models, amyloid formation is a multistep process involving the formation of collapsed oligomeric intermediates, their conversion to amyloidogenic nuclei, followed by polymerization of soluble material onto these nuclei. We used a conformation-specific antibody A11 [45] to detect molten oligomers that are on-pathway to Sup35 amyloid nucleation [15, 46]. All three variants accumulated A11-reactive species (Figure 4.11A). $Sup35^{Q}$ formed these species more rapidly than $Sup35^{WT}$

or $Sup35^{N}$, and remained in this form much longer. This suggests that the conversion of oligomeric intermediates into amyloidogenic nuclei is defective for $Sup35^{Q}$.

To address whether Ns and Qs also influence polymerization *per se*, we examined the rates at which freshly diluted soluble proteins polymerized onto their own preformed amyloid templates. Each variant was incubated with agitation for one week in assembly buffer and confirmed to have formed ThT-fluorescent, SDS-resistant aggregates (Figure 4.2F). These were sonicated into similar sized fragments and normalized to contain approximately the same number of fiber ends (Figure 4.12A-B). Nonlinear regression of ThT fluorescence kinetics was then used to determine initial polymerization rates across a range of added fiber concentrations (Figure 4.11B, 12C). The rates of seeded polymerization differed dramatically between variants. $Sup35^{N}$ converted more rapidly than $Sup35^{WT}$; $Sup35^{Q}$ converted much more slowly. The polymerization of Ure2 was altered in the same manner by Q substitutions (Figure 4.12A-D).

To discern if these results might simply be due to different oligomerization tendencies between Q and N variants, we also performed the complementary experiment. A range of concentrations of soluble proteins were seeded with a single concentration of preformed amyloids. A linear relationship was observed in all cases, indicating that the amyloids for all variants elongated by monomer addition under these conditions (Figure 4.12F-G).

Next, preformed sonicated amyloids of each of the Sup35 and Ure2 PrD variants were used to cross-seed amyloid formation by each of the other variants. In all but one case, cross-seeding was not observed, indicating that Ns and Qs generally create incompatible templates (Figure 4.11C, 12E). The single exception occurred between the

pair of proteins with the greatest sequence identity: Sup35^{WT} and Sup35^{Q}. This relationship was asymmetric: Soluble Sup35^{Q} protein did not polymerize onto Sup35^{WT} amyloid. However, soluble Sup35^{WT} effectively polymerized onto Sup35^{Q} amyloids. Thus, Sup35^{Q} amyloids are not defective for templating *per se*. Rather, the reduced polymerization of Sup35^{Q} appears to derive from a relative inability of its non-amyloid conformers to productively engage with amyloid templates.

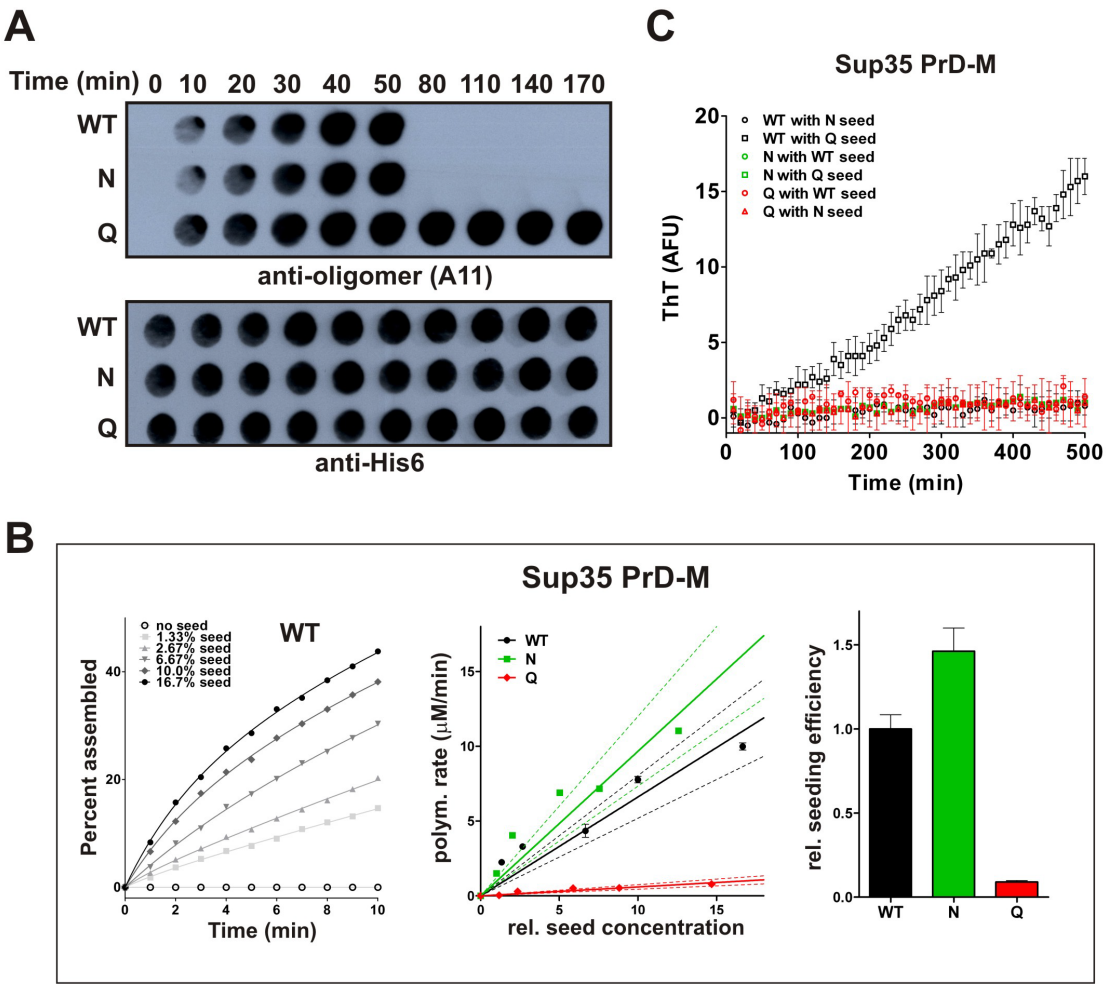
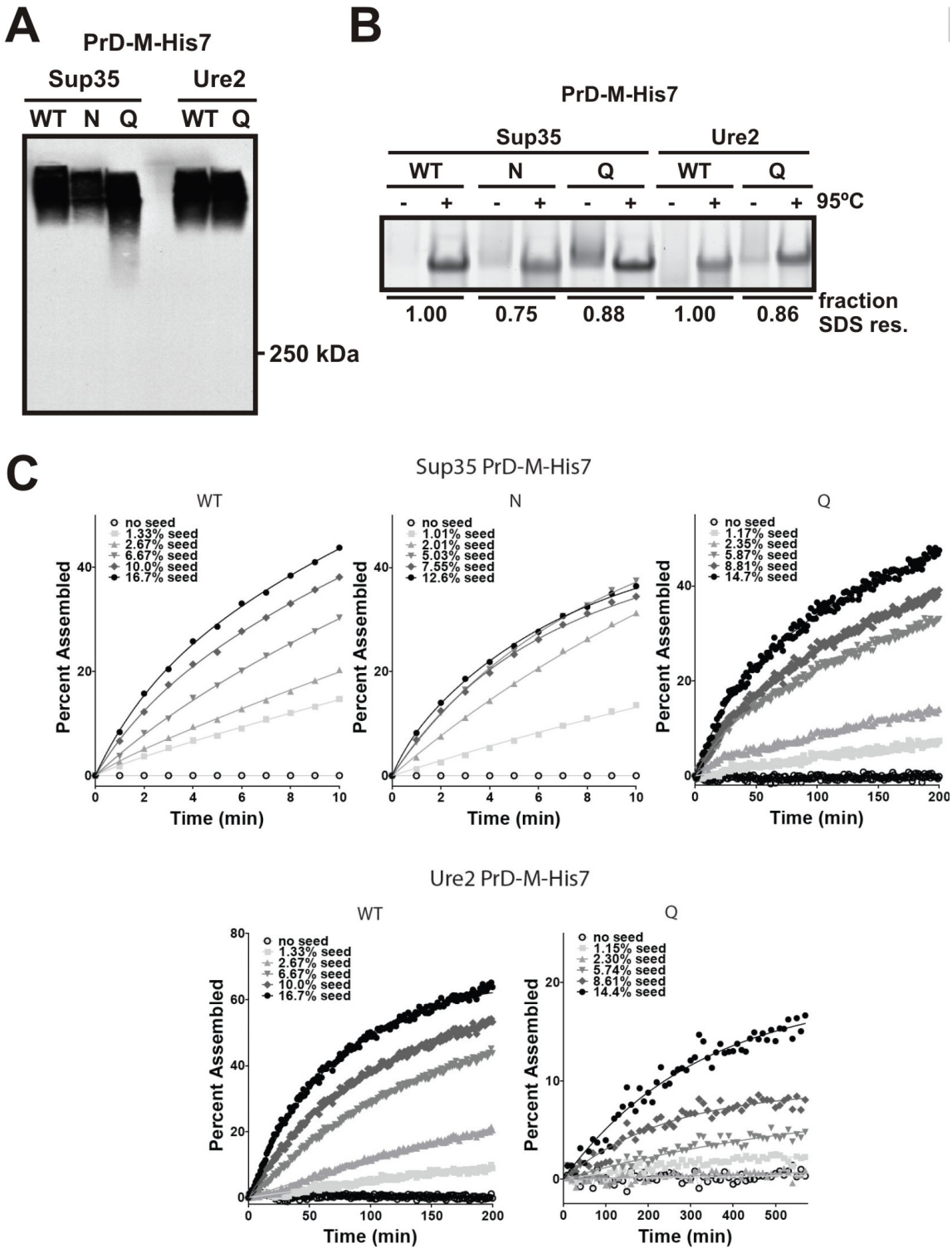


Figure 4.11: Q-rich proteins have reduced rates of conformational conversion to amyloid. (A) Sup35 PrD-M-His7 variants were diluted to 2.5 μM in assembly buffer and

incubated for the indicated times prior to the removal of 50 µl to a nitrocellulose membrane. Pre-amyloid oligomers (top) or total protein (bottom) were detected with A11 or anti-His6 antibodies respectively. (B) Sup35 PrD-M-His7 variants were diluted to 7.5 µM in assembly buffer containing ThT, followed immediately by the addition of various concentrations (% m/m) of the respective preformed sonicated amyloid fibers. Reactions were incubated without agitation and monitored for amyloid polymerization by ThT fluorescence. Nonlinear regression (as shown on left for WT, fit to one-phase association curves) was used to determine initial rates of amyloid elongation (as shown in middle, plotted against normalized seed concentrations). Dotted lines denote the 95% CI of the best-fit line. Slopes of the best-fit lines show the seeding efficiencies of each variant amyloid preparation, relative to WT (right). (C) The ability of individual variants to polymerize onto heterologous pre-assembled amyloids. 5 µM soluble protein was seeded with 10% (m/m) preformed aggregates in each case. Data show means +/- SEM.


A
PrD-M-His7
Sup35
Ure2
WT N Q WT Q
250 kDa
B
PrD-M-His7
Sup35
Ure2
WT N Q WT Q
- + - + - + - + - + 95ºC
1.00 0.75 0.88 1.00 0.86
fraction SDS res.
C
Sup35 PrD-M-His7
WT
no seed
1.33% seed
2.67% seed
6.67% seed
10.0% seed
16.7% seed
N
no seed
1.01% seed
2.01% seed
5.03% seed
7.55% seed
12.6% seed
Q
no seed
1.17% seed
2.35% seed
5.87% seed
8.81% seed
14.7% seed
Percent Assembled
Time (min)
Ure2 PrD-M-His7
WT
no seed
1.33% seed
2.67% seed
6.67% seed
10.0% seed
16.7% seed
Q
no seed
1.15% seed
2.30% seed
5.74% seed
8.61% seed
14.4% seed
Percent Assembled
Time (min)

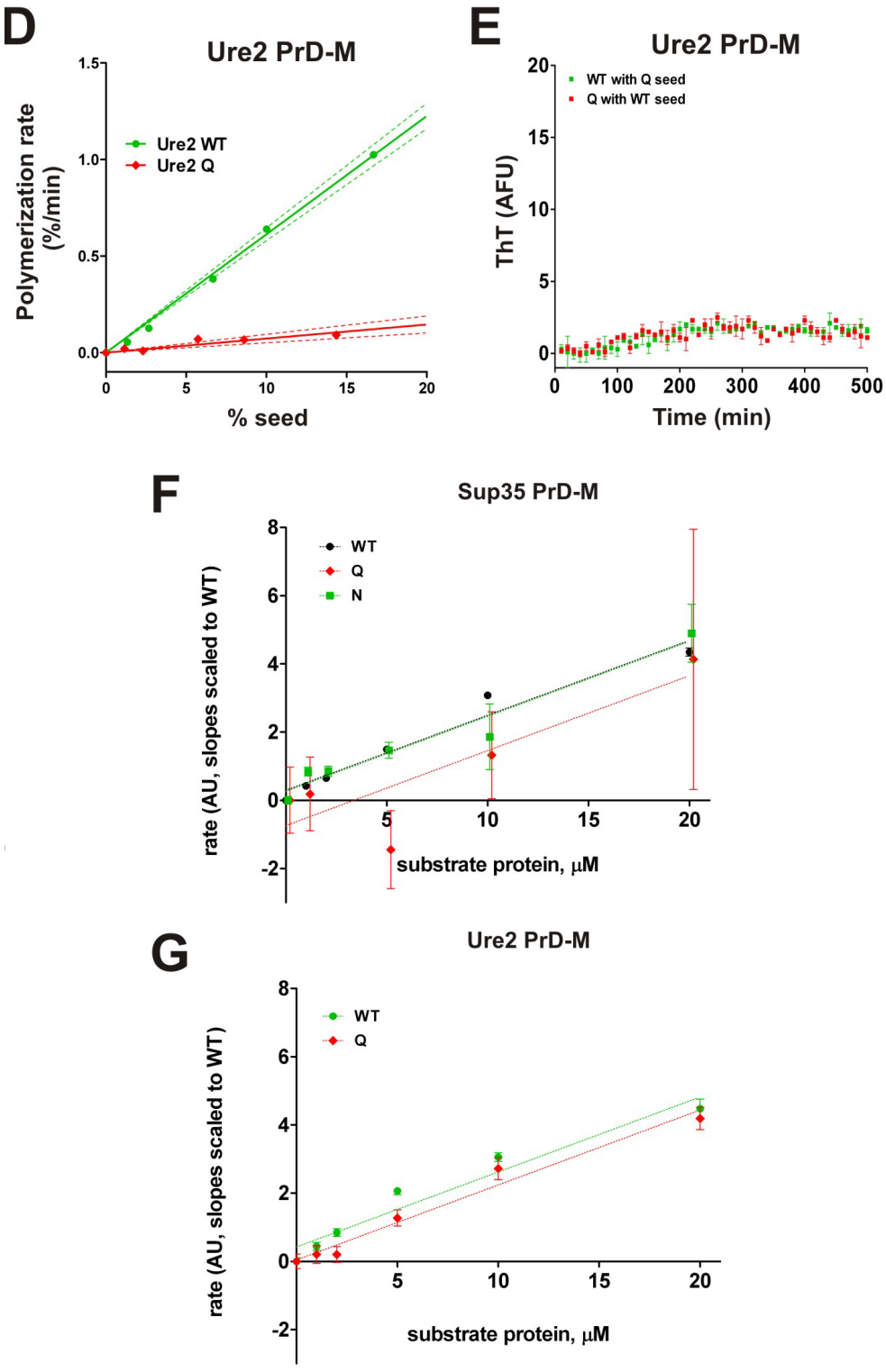

D
Ure2 PrD-M
Polymerization rate (%/min)
Ure2 WT
Ure2 Q
% seed
E
Ure2 PrD-M
WT with Q seed
Q with WT seed
ThT (AFU)
Time (min)
F
Sup35 PrD-M
WT
Q
N
rate (AU, slopes scaled to WT)
substrate protein, μM
G
Ure2 PrD-M
WT
Q
rate (AU, slopes scaled to WT)
substrate protein, μM

Figure 4.12: related to Figure 4.11. (A) Aggregates of Sup35 and Ure2 PrD-M-His7 variants assembled as in Figure 4.2F were collected by ultracentrifugation, resuspended in fresh assembly buffer, sonicated, and then examined by SDD-AGE. This ensured that the preparations contained approximately the same number of fiber ends per mass of polymer. Note that SDD-AGE often fails to detect protein monomers. Consequently, the absence of detectable monomers on this blot does not indicate an absence of SDS-soluble species in the assemblies. (B) SDS-resistant protein was quantified by differential entry into an SDS-PAGE gel following incubation in sample buffer at either 25°C or 95°C. (C) The indicated proteins were diluted to 7.5 μM in assembly buffer containing the indicated concentrations of SDS-resistant pre-assembled fibers, and monitored for ThT fluorescence over time. ThT data were normalized by the calculated fluorescence for 100% assembly (final ThT data for 7.5 μM seed stocks, divided by SDS-resistant fraction). Data from three replicates each were fit to one-phase association curves. To achieve a better fit at early time points, $Sup35^{Q}$ was fit to two-phases. Data indicate means +/- SEM. (D) As in Figure 4.11B middle, for Ure2 PrD-M-His7 variants. (E) As in Figure 4.11C, for Ure2 PrD-M-His7 variants. (F-G) Dependence of polymerization rates on substrate concentration. Purified denatured proteins were diluted to the indicated concentrations in assembly buffer containing a fixed concentration of seed (0.02 μM total seed stock for $Sup35^{WT}$, $Sup35^{N}$ and $Ure2^{WT}$, or 0.2 μM total seed stock for $Sup35^{Q}$ and $Ure2^{Q}$) and examined for ThT fluorescence over time. Initial assembly rates were determined as in Figure 4.12C. For qualitative comparisons of the relationship between assembly rate and substrate concentration, data for each variant were normalized by the

slope of the linear regression against added substrate protein. Data indicate means +/- SEM.

4.8 Towards a mechanistic distinction between Qs and Ns

Why does a subtle chemical distinction between N and Q side chains, namely, one methylene group, so strongly influence amyloid propensity? The conformational fluctuations that lead a disordered protein to convert to amyloid are difficult to dissect experimentally. Molecular simulations provide a complementary tool for investigating the free energy landscapes and thermodynamics of β-sheet formation in such sequences [(47-51)]. To understand on the intrinsic differences between Q- and N-rich sequences, without the confounding complexities imposed by different sequence contexts, we performed molecular simulations with polyQ and polyN molecules. For practical reasons we limited the simulations to molecules containing 30 glutamines (Q_{30}) or 30 asparagines (N_{30}). We performed two sets of simulations. In one set, we interrogate the unbiased free energy landscapes and in the second set, we imposed local conformational restraints to generate non-specific biases of the backbone dihedral angles in the β-basin of conformational space. The latter set allows us to observe rare conformations that might be sampled on-pathway to amyloid formation.

Conformational restraints allow us to design simulations where the entropic penalty is pre-paid in equivalent fashion for both Q_{30} and N_{30}. Our analysis focused on two quantities viz., the degree of ordered intramolecular β-sheet formation in the presence of conformational restraints and the probability that a pair of conformationally biased molecules would self-associate. Figure 4.13A compares the degree of ordered β-sheet formation in N_{30} and Q_{30} in the presence and absence of conformational restraints.

This was quantified using DSSP-E scores (Figure 4.13A) whereby higher scores imply greater degree of regular hydrogen bonding characteristic of canonical β-sheets. For monomeric N_{30} and Q_{30}, the extent of β-sheet formation is low and equivalent in the absence of conformational restraints. Conversely, monomeric forms of N_{30} show increased ordered β-sheet formation when the entropic penalties for sampling the appropriate conformations have been pre-paid. The effects of homotypic intermolecular interactions were simulated using two N_{30} molecules. Intermolecular interactions neither diminished nor enhanced the intrinsic β-sheet-propensity. Monomeric Q_{30} showed greatly reduced ordered β-sheet content even in the presence of biases that restrict the backbone dihedral angles to the β-basin. However, in simulations with two restrained Q_{30} molecules, there was positive coupling and the overall β-sheet content of both Q_{30} molecules increased through intermolecular interactions, suggesting that ordered β-sheet formation in Q-rich systems requires at least two interacting molecules [(52)] that have been appropriately biased to sample conformations drawn from the β-basin.

Next, we quantified the thermodynamics of bimolecular associations (Figure 4.13B). The probability of intermolecular associations was smaller for N_{30} than for Q_{30}. The intermolecular associations in such simulations are largely non-specific [(47, 48)], i.e. spontaneous fluctuations lead disordered monomers to form disordered dimers. The presence of conformational restraints decreased this disorder and, in turn, caused systematic diminutions in the magnitudes of intermolecular associations, an observation borne out by the temperature-dependence of these probabilities. The lower disorder for N_{30} and its increased ability to form ordered β-sheet structures (Figure 4.13A), leads to weaker non-specific intermolecular associations. This analysis makes the point that

conformational heterogeneity is a pre-requisite for favorable non-specific associations. Q_{30} molecules show preference for increased conformational heterogeneity and hence increased preference for non-specific associations.

The different non-specific association tendencies of Q_{30} and N_{30} appeared to result, at least in part, from a difference in turn formation between the two systems (Figure 4.13C). No more than four Ns were needed to form a tight turn. These were often canonical β-hairpin turns (Figure 4.13C) with characteristic intra-turn distances, backbone dihedral angles, and hydrogen bonding patterns. Conversely, the bulkier side chain of Q in Q-rich tracts cannot form tight turns but instead forms wider bulges and loops that require at least five (often more) residues to promote the reversal of chain direction.

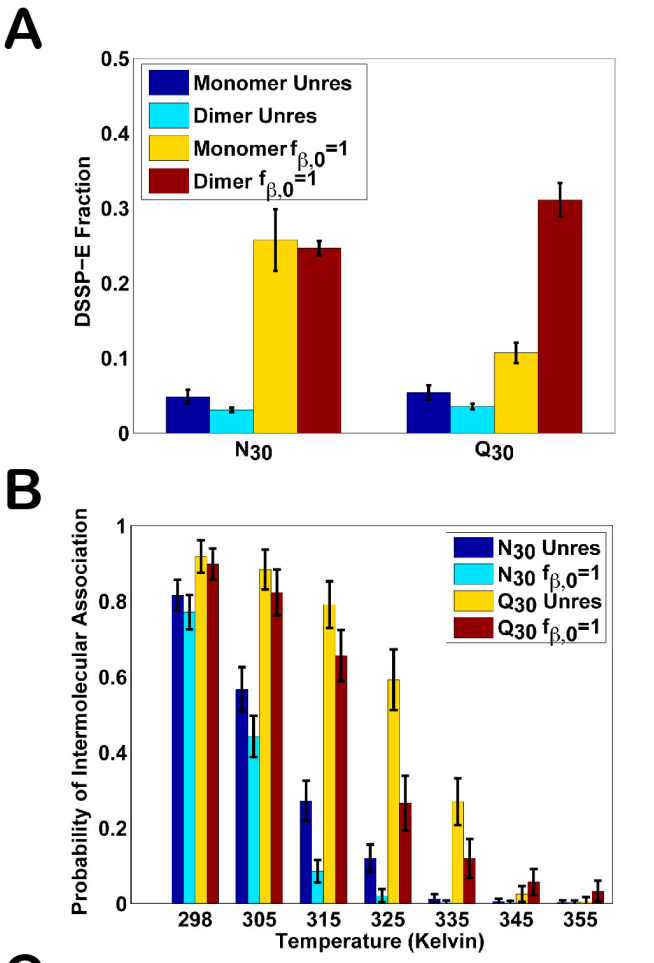

A
Monomer Unres
Dimer Unres
Monomer $f_{\beta,0}$=1
Dimer $f_{\beta,0}$=1
DSSP-E Fraction
0.5
0.4
0.3
0.2
0.1
0
N30
Q30
B
N30 Unres
N30 $f_{\beta,0}$=1
Q30 Unres
Q30 $f_{\beta,0}$=1
Probability of Intermolecular Association
1
0.8
0.6
0.4
0.2
0
298
305
315
325
335
345
355
Temperature (Kelvin)
C
N30
Q30
N21
N24
N22
N23
Q13
Q18
Q17
Q14
Q16
Q15

Figure 4.13: Molecular simulations of polyN (N_{30}) and polyQ (Q_{30}). (A) Percentage of ordered β-sheet formed by N_{30} and Q_{30}. Single N_{30} and Q_{30} molecules were simulated in the absence (dark blue) or presence (yellow) of local conformational restraints that restrict conformational sampling to dihedral angles drawn from the β-basin in conformational space. Pairs of N_{30} and Q_{30} molecules simulated with (cyan) or without (dark brown) local conformational restraints show the effects of homotypic intermolecular interactions on ordered β-sheet content. (B) Temperature-dependent probabilities of realizing homotypic intermolecular associations, quantified as the probability that the intermolecular (center-of-mass to center-of-mass) distance between the pair of restrained / unrestrained N_{30} or Q_{30} molecules is ≤ 25Å (corresponding to less than 0.025% of the total volume available to the molecules in the simulation setup). Simulations were performed for pairs of N_{30} and Q_{30} molecules without (dark blue and yellow) and with (cyan and dark brown) local conformational restraints. (C) Visual comparison of ordered β-sheet structures formed by N_{30} (left) and Q_{30} (right) molecules in the presence of local conformational restraints. Note the tight type I β- turn formed by N_{30} relative to Q_{30}, and the resulting differences in the lengths of intramolecular antiparallel β-sheets. When the entropic penalty is pre-paid using conformational restraints, we find a greater frequency of sampling intramolecular β-sheet structures with N_{30} because asparagine tracts can form canonical β-turns through backbone and sidechain hydrogen bonds, a representative of which is shown in the enlarged picture in green for N_{30}. Conversely, Q-rich tracts form longer loops that lack any of the hallmarks of canonical turns and this increases the barrier for strand nucleation and propagation [(53)].

4.9 Proline containing turn motifs in Sup35 enhance turn formation

Observations by Alberti et al. and Toombs et al. suggest the presence and sequence patterning of proline residues are relevant to prion formation. Within a beta strand, proline acts as a structure breaker and can be inhibitory to prion formation. Alternatively, prolyl residues placed in the appropriate context can promote turn formation, which can help nucleate β–hairpin motifs. In fact, proline residues are present in several well-known β–turn motifs. It is possible that correctly placed proline residues can facilitate prion formation by the same mechanism that we propose N-rich sequences do: by lowering the barrier to β–hairpin formation. The kinetics of amyloid formation for Sup35 (Figure 4.1F), Ure2 and Lsm4 (Figure 4.3D) provide evidence in support of this hypothesis. N-rich mutants of Ure2 and Lsm4 are enhanced in amyloid formation vis-à-vis their Q-rich counterparts. Wild type Sup35 contains four putative turn forming motifs (a single NPDA and three NPQG motifs), each containing a single proline spaced roughly 9 residues apart. The presence of these motifs is likely relevant to observed enhancement in the rate of amyloid formation with respect to the N-rich Ure2 and Lsm4 mutants. We propose that N-richness and appropriately placed proline residues work in concert to enhance β–hairpin formation and in turn they increase the amyloidogenicity of the host protein.

We studied turn formation by NPDA and NPQG using molecular simulations and analysis of statistics from the protein data bank (PDB). Molecular simulations were conducted on these motifs. In these simulations we also include three N- and C-terminally flanking residues from wild-type Sup35 to allow for turn stability. The two

turn sequences are as follows: QQYNPQGGYQ and QQYNPDAGYQ. For both sequences, mutations of all Q's to N's and all N's to Q's were also examined.

The PDB statistics in Figure 4.14 and 4.15 show that both motifs form β-turns in proteins of known structure. These turns are characterized by turn dihedral angles of 0 to 100 degrees and α-carbon i+1 to i+4 distances less than 7 Å. These geometric descriptors are the same as those for the simulated N_{30} constructs. Note that in the case of Q_{30} constructs, α-carbon i+1 to i+4 distances were almost always greater than 7 Å. The simulated turn motifs form β-turns as evidenced by the high probability for the ensemble to occupy the same turn dihedral angles and α-carbon i+1 to i+4 distances as those found in the PDB. In the absence of full length Sup35, it is reasonable that these short peptides do not always remain turns as a smaller fraction of the ensemble forms coil conformations. Access to both conformational states suggests that our simulation model is not overly biased for either turn or coil formation. The ability of these motifs to form turns is not significantly altered by the N- or Q-richness of the flanking sequence suggesting that the effect from the presence of the proline turn motif is not abrogated by Q-rich sequences and shows that our model does not unreasonably bias against turn formation for Q-rich sequences or over bias for N-rich sequences. Both the NPQG and NPDA motifs promote the formation of β-hairpin turns. Turn formation is not obliterated by an overabundance of Q's. This is consistent with the observation that $Sup35^{WT}$ can form amyloids, albeit at a slower rate than $Sup35^{N}$. In all Sup35 variants, the putative turn motifs are in the right location. However, the significant diminution of amyloid formation in $Sup35^{Q}$ suggests that the presence of prolyl residues and / or the turn motifs alone is insufficient to guarantee conversion to amyloid. In the presence of the

appropriate turn motifs it is also necessary to achieve a proper balance of Qs and Ns (leaning toward more Ns) thus further corroborating the importance of the intrinsic differences between Qs and Ns identified in this study. The preliminary simulations of turn forming propensities of NPDA and NPQG coupled to simulations of polyQ and polyN suggest that the extension of these simulation investigations for comparative analysis of N- and Q-rich variants of Sup35 might generate useful, predictive insights regarding the monomer ensembles and mechanism of self-assembly.

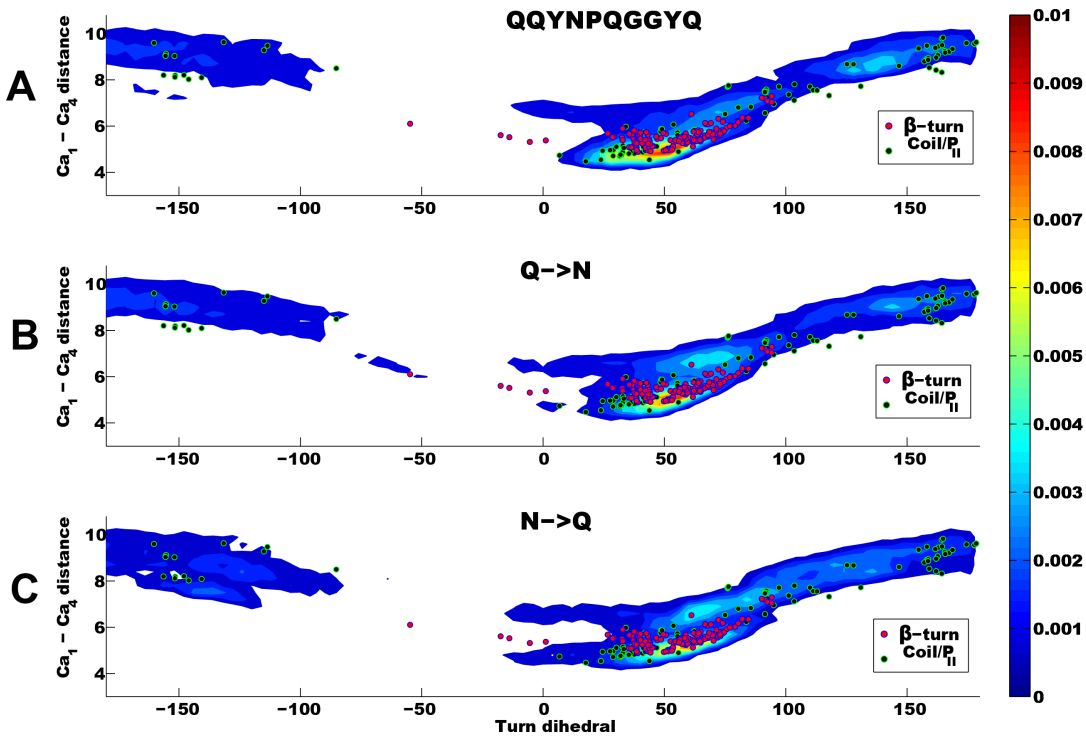


Figure 4.14: Joint histograms of NPQG turn dihedral angles and α-carbon i+1 to i+4 (N to G) distances. The turn dihedral angles and N to G α-carbon distances in angstroms were measured for all occurrences of the NPQG motif in the PDB and are shown as dots overlaid on the joint histogram. The PDB structural data was annotated by the PROSS *(54)*

secondary structure assignment tool to distinguish NPQG motifs that form β-hairpin turns (pink dots with black outline) from those that are generic coils (black dots with green outline) and this data is repeated in each panel. The simulation data was accumulated every 5000 steps and is shown in each panel for peptides (A) QQYNPQGGYQ, (B) NNYNPNGGYN, and (C) QQYQPQGGYQ using a bin size of 0.3Å and 4°. The color bar shows the probability of occurrence for each distance-dihedral combination.

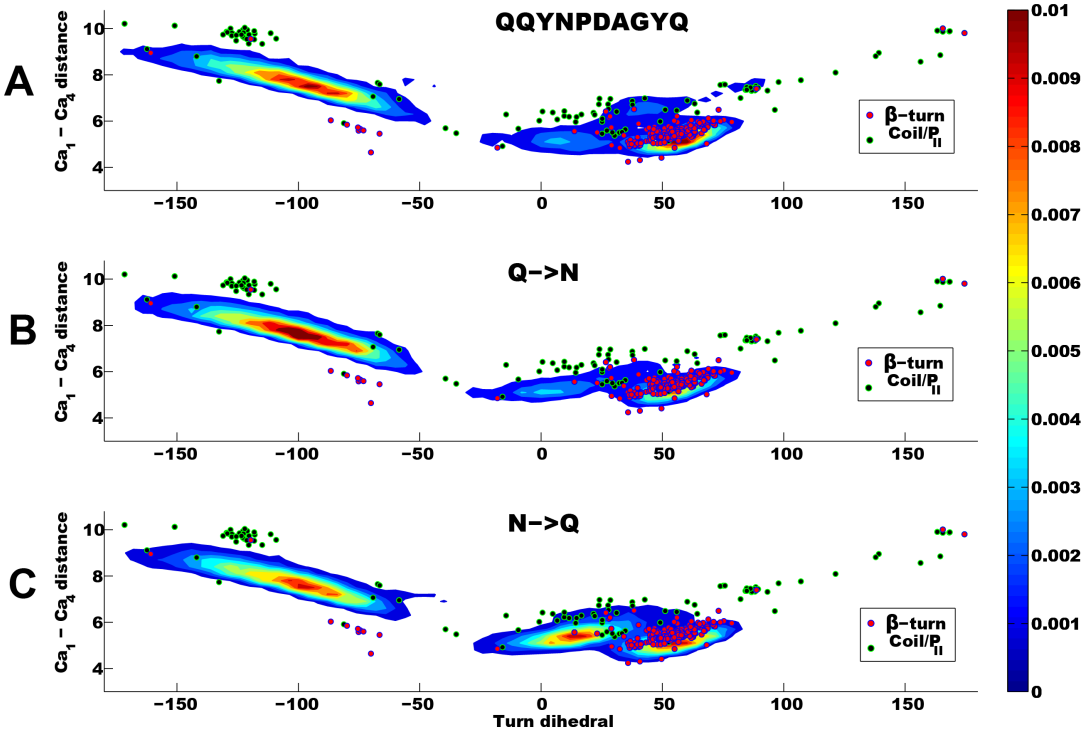


Figure 4.15: Joint histograms of NPDA turn dihedral angles and α-carbon i+1 to i+4 (N to G) distances. This is analogous to the analysis in Figure 4.14 except over the turn motif NPDA. Simulated peptides were (A) QQYNPDAGYQ, (B) NNYNPDAGYN, and (C) QQYQPDAGYQ.

4.10 Discussion

In recent years proteins with intrinsically disordered regions have engendered considerable interest. Such regions can have tethering or bristle-like functions in extended states, but more often drive the binding interactions that govern hubs in regulatory networks [3, 55]. Disordered proteins also feature prominently in protein misfolding diseases [3]. We have focused on a subset of disordered proteins with the capacity to form protein-based genetic elements in yeast. Analyzing the defining sequence feature of this subset, the Q/N-rich "prion-like" proteins, we find that Qs and Ns strongly and disparately affect transitions into higher order complexes. In diverse Q/N-rich proteins, changing Qs to Ns greatly enhanced amyloid- and prion-forming capabilities. Changing Ns to Qs had a different effect, resulting in the accumulation of non-amyloid assemblies that were toxic in both yeast and mammalian systems. Q and N content affected not only the ability of oligomers to mature into amyloids, but also the efficiency with which soluble proteins could be templated by amyloids, once formed.

Our molecular simulations provide a preliminary rationalization for the disparities between Q-rich and N-rich sequences. The shorter N side chain can enhance the ability of this residue to hydrogen bond to the polypeptide backbone. This leads to easier turn-formation, increased β-sheet propensity, and a decreased tendency for nonspecific intermolecular associations, providing the entropic penalties for sampling β-like conformations is pre-paid. We propose that these distinctions become amplified as multiple monomers come together and, as a result, N-rich molecules more effectively form ordered self-assemblies on pathway to amyloid. Regardless of the mechanism of amyloid nucleation – homogeneous monomeric nucleation [56], or nucleated

conformational conversion within molten oligomers [15, 18] – the barrier for molecular conversion to an ordered, β-sheet-rich conformation appears to be lower for N-rich sequences. Furthermore, by reducing nonspecific interactions, N-richness is likely to lessen the depletion of soluble species through off-pathway aggregation, which competes with amyloid polymerization [57].

We acknowledge a wide epistemological gap between molecular behaviors amenable to computer simulations and the macroscopic polymerization and phenotypic behaviors of Q/N-rich proteins. Clearly, further experimentation will be required to test the proposed relevance of Q and N turn-forming tendencies to amyloid formation. Recently, Fiumara et al. reported that many Q/N-rich proteins contain coiled coil regions that overlap their amyloidogenic regions, and indicated that coiled coil formation enhances aggregation. However, N has a lower coiled coil propensity than Q (Paircoil2 [58]: 0.29 vs. 0.90; Coils [59]: 0.25 vs. 0.99; for polyN and polyQ, respectively). Thus, if coiled coil formation contributes to the distinct prion propensities of Ns and Qs, it is not in the manner predicted from Lupas et al. Instead, the prion propensities of our Q/N-substitution variants correlate negatively with coiled coil propensities.

Likewise, we cannot formally exclude the possibility that the differences we observe result from substitutions in only a few crucial positions, among the many made in each variant. For instance, a minimal content of N may be necessary for turn formation, with additional Ns having little effect. Our global Q and N substitutions cannot discern the existence of such a threshold. We note that Ns, but not Qs, are commonly involved in hydrogen-bonded spines, or “asparagine-ladders”, of β-helices in

the Protein Data Bank [60]. The β-helix is a model structure for several functional amyloids, including those of the prions formed by Het-s [61] and Sup35 [19, 62, 63].

The disparity we observed between Qs and Ns was surprising. The notion that Q and N are interchangeable for prion formation is pervasive. Many sequences in the prion literature are described as "Q/N-rich" though they are, in fact, often enriched for only one of these residues [64-68]. Algorithms commonly used to identify amyloidogenic sequences – TANGO [69] and Zyggregator [70] – failed to yield a clear prediction about the effects of Q and N replacements (Figure 4.16). They also failed to predict amyloid formation by Q/N-rich PrDs [11] or the PrD variants analyzed in this work (Figure 4.16). The distinctions revealed here should begin to provide information for improved sequence-based predictions of amyloid formation.

Amyloids have a wide range of structural functions, ranging from peptide hormone storage and biopolymer synthesis in mammals, to spore dispersal and cellular adhesion in microbes [71, 72]. Biofilm-forming amyloids driven by N-rich sequences have recently been identified in *Pseudomonas* [21]. In yeast and perhaps in other organisms, N-rich amyloids also function as protein-based elements of inheritance – prions. The stochastic switching of prion proteins to amyloid states, and the stochastic loss of those states, may be important for maintaining adaptive phenotypic diversity within clonal cell populations [27, 28, 32]. Many N-rich sequences are found in roundworm and insect proteomes, and they are also abundant in certain lower eukaryotes such as *Dictyostelium* and *Plasmodium* [29, 73, 74]. Based on our findings it is tempting to suggest that the biological functions of these proteins involve self-assembly into amyloid.

While amyloids represent one extreme of conformational space (highly ordered and stable) many functions of Q/N rich proteins derive from the other conformational extreme they populate namely, intrinsic disorder. Qs and Ns are predicted to have roughly equivalent disorder-promoting tendencies [8] although other disorder predictors suggest an increase of Qs within IDRs vis-à-vis Ns. Our Q and N variants are all predicted to be highly disordered (9 of 10 disorder prediction web-servers reviewed in [75]). Yet, intrinsically disordered proteins (aside from prions) are typically enriched for Qs and depleted of Ns [2]. This bias is most prevalent in the proteomes of mammals, which contain many more Q-rich sequences than N-rich sequences [29, 73, 76-78]. It cannot be attributed to differences in codon frequencies or to structural properties of the DNA, but instead appears to reflect positive selection pressures acting on Q-rich proteins [78, 79]. Our findings suggest that the puzzling deficiency in Ns results from selective pressure against amyloid formation, which would inactivate the functions of these proteins that depend on disorder. In agreement, the bias against Ns in diverse proteomes increases with the length of the disordered region [80]. Longer disordered regions have an increased risk for amyloid formation [9, 56, 81], making them particularly susceptible to the amyloidogenicity imparted by N-richness.

Molecular simulations also suggest that the protein:protein interactions of Q-rich proteins are highly unstructured [47, 48, 51]. This property is likely integral to the functions of Q-rich proteins in many large and dynamic protein assemblies: transcriptional regulatory complexes, RNA processing bodies and endocytic complexes [4, 11, 67, 82-85]. We suggest that the conformational heterogeneity of Q-rich polypeptides expedites the assembly and remodeling of these complexes. Further, that Q-rich protein interactions are

structurally less constrained may grant the freedom to explore new binding partners, accelerating the functional diversification of network hubs and the evolution of novel circuitries.

But there can be a price to be paid for these properties. The conformational disorder of Q-rich sequences increases their burden on protein homeostasis. Disordered proteins tend to be toxic when over-expressed, in part due to mass-action driven interaction promiscuity [(5)]. This liability may drive the extraordinarily tight regulation of the expression of intrinsically disordered proteins in general, and "Q/N-rich" proteins in particular [(6)]. Q-richness appears to increase the propensity for toxic interactions by disordered proteins (Figure 4.8E), which, in turn, may contribute to the pathology of Q-rich proteins in disease. We find that amyloid formation reduces the toxicity of over-expressed Q-rich proteins, presumably by sequestering the protein species prone to making toxic associations. These observations add to a growing body of evidence from a wide range of proteins and disease models that amyloids often reduce, rather than exacerbate, the consequences of protein misfolding [(86-88)].

We also find that Q-richness causes proteins to dwell in an oligomeric state that is reactive with the conformation-specific antibody, A11. Such oligomers, therefore, must share conformational features with oligomeric intermediates of Aβ, against which the antibody was raised, despite the fact that Aβ has only a single Q residue in its 42 amino acid sequence. In work from Krishnan, et al. [(89)], several other manipulations of the Sup35 PrD that cause it to dwell in an A11-reactive oligomeric state also lead to toxicity, which was found to be greatest when oligomer-rich preparations were applied to cells extracellularly. Of course, comparisons will need to be made between proteins expressed

in, and applied to, a variety of cell types and compartments, but it may be that the toxicity of such species derives from multi-factorial effects that not only involve intracellular protein-protein interaction promiscuity but also membrane permeabilizing activities. The genetic tractability of yeast prions, and the fact that they are normally benign, provides a new tool for investigating this very difficult problem.

A single methylene distinguishes Q from N. Yet this difference unequivocally alters one prominent activity of Q/N-rich proteins, prion formation, and also influences another, toxicity. The presence and placement of turn forming motifs are an additional modulator of these systems. Further understanding the conformational preferences of disordered proteins will be key to elucidating their widespread roles in both normal biology and disease.

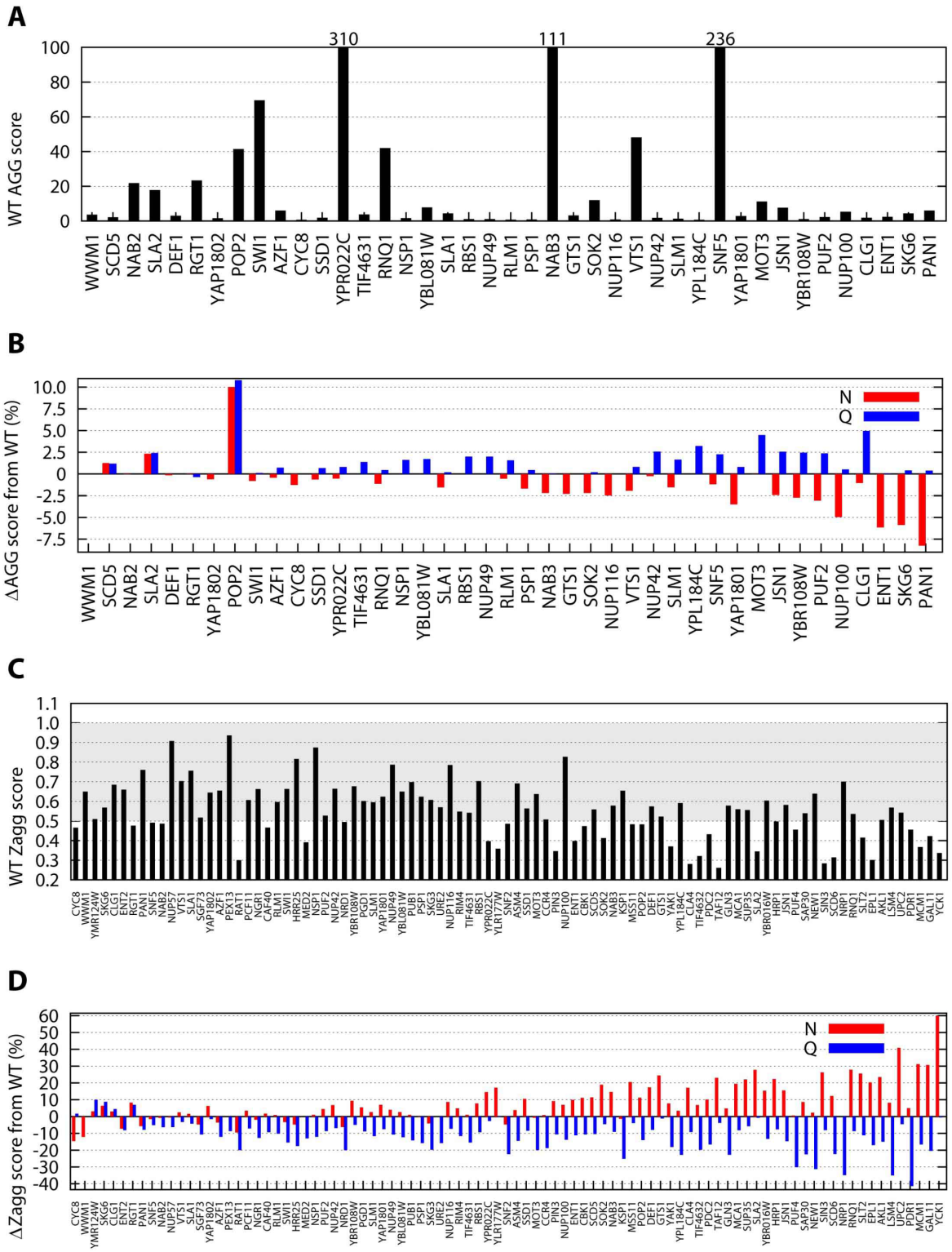


Figure 4.16: Amyloid prediction algorithms do not predict a strong difference between N and Q. TANGO and Zyggregator predictions were run on all 92 Q/N-rich

sequences from Alberti et al. [11] using a pH of 7. For each sequence, N→Q and Q→N predictions were also performed. (A) TANGO predicted an *AGG* score of 0 for 51 of the 92 sequences. Shown are sequences with non-zero scores. Such low scores generally indicate very low likelihoods for amyloid formation; for comparison, the $A\beta_{1-42}$ peptide received a score of 1565. (B) For the 41 sequences receiving a non-zero *AGG* score, the percent change in *AGG* score between WT and N→Q substitutions and between WT and Q→N substitutions is given. In 39 cases, the N→Q substitution produced higher amyloid propensity. (C) Zyggregator scores (*Zagg*) for all 92 sequences are provided. *Zagg* score values range between 0.5 and 1.0 for most peptides and proteins (shaded region). Scores below 0.5 suggest unusual resistance to aggregation while scores above 1.0 are considered aggregation-prone [90]. All scores fall within the normal or aggregation-resistant range. (D) The percent change in *Zagg* score between WT and N→Q substitutions and between WT and Q→N substitutions is given.

4.11 Conclusion

Sequences rich in glutamine and asparagine residues often fail to fold at the monomer level. This coupled to their hydrogen-bonding ability provides the driving force to switch between disordered monomers and amyloids. Such transitions influence processes as diverse as human protein-folding diseases, assembly of bacterial biofilms and formation of protein-based genetic elements in yeast (prions). A systematic survey for prion-forming domains suggested that Q and N residues have distinct effects on amyloid formation. We used cell biological, biochemical, and computational techniques to compare Q/N-rich protein variants, wherein Ns were replaced with Qs or Qs with Ns. Qs and Ns had strong and opposing effects: N-richness promotes assembly of benign

self-templating amyloids; Q-richness promotes formation of toxic non-amyloid conformers. Molecular simulations that focus on the intrinsic differences between Qs and Ns suggest that the observed effects can be attributed to the enhanced turn-forming propensity of Ns over Qs. Simulations also show two sequences NPQG and NPDA can promote turn formation and act in concert with Q/N content to modulate the phase behavior of these disordered proteins.

4.12 Materials and methods

4.12.1 Cloning and gene synthesis

Cloning procedures were essentially performed as described previously (Alberti et al., 2009). Variant versions of PrDs and Huntingtin exon 1 were synthesized and assembled by DNA2.0 (Menlo Park, CA) and then cloned into the pDONR221 plasmid. The coding sequences were codon-optimized for expression in yeast and contained flanking sequences that allowed for Gateway® recombination and dual expression in yeast and bacteria [(11)]. Additional entry clones were generated for the PrDs (lacking the M domain) of each Sup35 variant. PCR reactions used Platinum Pfx DNA polymerase (Invitrogen, CA), variant Sup35 PrD-M entry clones as templates. The correct amplification and integration of DNA into pDONR221 (Invitrogen, CA) was confirmed by sequencing. ORFs in entry clone format were transferred into the following destination vectors: pAG424GAL-ccdB-EYFP [(11)], pAG415SUP35-ccdB-SUP35C, pAG415ADH1-ccdB-SUP35C, pAG415GPD-ccdB-SUP35, pRH1 and pRH2 [(11)].

Note that our Htt constructs differ from those previously used to study toxicity of polyQ in yeast [(40)], in that ours do not contain a FLAG tag. The FLAG tag causes polyQ-

expanded Htt to become highly toxic in [RNQ+] yeast [91]. We purposefully avoided this scenario, as it would have complicated our analyses of amyloid formation.

4.12.2 Yeast techniques

Standard genetic manipulations, media conditions, and fluorescence microscopy were as described [11]. Sup35 variants for prion maintenance, prion induction, and microscopy consisted of the entire prion-determining region of Sup35 (PrD and M domains), fused to either Sup35C or EYFP. All other yeast experiments and protein variants utilized PrD regions only. For cell spotting assays (prion induction and toxicity), PrD-EYFP fusions were expressed from a high copy galactose inducible plasmid. For prion induction, cells were grown overnight in galactose- prior to plating on glucose-containing media. For toxicity, cells were grown overnight in glucose- prior to plating on either galactose- or glucose-containing media. Plates were incubated at 30°C. For toxicity in liquid media, cells were grown overnight in raffinose media (non-repressing but not inducing) and then diluted to OD 0.05 in galactose media. Cells were then shaken at high speed on a BioscreenC instrument (Oy Growthcurves, Ab Ltd. Helsinki, Finland), with growth monitored by absorbance at 600nm. Additional details are provided in the supplement.

4.12.3 SDD-AGE

SDD-AGE was performed as described [11].

4.12.4 Protein purification

All proteins were expressed and purified from *E. coli* BL21-AI essentially as described [11], using either pRH1 (for fusing a 7xHis tag to the C-termini of Sup35 PrD-M, Ure2 PrD, and Lsm4 PrD variants) or pRH2 (for fusing a Sup35 M domain plus 7xHis

tag to the C-termini of Ure2 PrD variants). Sup35 variants were further purified by seeded polymerization in assembly buffer, with rotation, for one week, followed by recovery of aggregated protein by ultracentrifugation for one hour at 100,000 rcf. Truncation products and other co-purified contaminants remained in the supernatant. The pellet was re-dissolved in 6 M GdnHCl followed by precipitation with 5 volumes methanol at -80°C. Methanol-precipitated proteins were resuspended in 6 M GdnHCl, incubated for 5 min at 95°C, and then filtered through a YM-100 Microcon filter immediately prior to use.

4.12.5 *In vitro* aggregation assays

For reactions monitoring the rate of amyloid formation, proteins were diluted into assembly buffer (5 mM K_2HPO_4, pH 6.6; 150 mM NaCl; 5 mM EDTA; 2 mM TCEP) plus 0.5 mM ThT, in black nonbinding microplates (Corning, NY) with 100 µl per well. GdnHCl concentrations in reactions did not exceed 60 mM, and were equalized in all pairwise comparisons. *De novo* amyloid assembly reactions monitored by ThT fluorescence were incubated at 25°C and shaken 10 sec every 2 min. For unseeded reactions with Sup35 variants, 3 PTFE 3/32" plastic beads (McMaster-Carr) were added to each well to increase the rate of assembly. Seeded reactions were not shaken. Fluorescence measurements (450 nm excitation, 482 nm emission) were made with a Sapphire II plate reader (Tecan, NC). For seeded reactions, data were fit to one phase (pseudo-first order) associations using GraphPad Prizm software. To achieve a better fit at early time points, $Sup35^Q$ was fit to two phases. For experiments requiring larger reaction volumes (oligomer formation kinetics, preparation of aggregates for fractionation and membrane disruption assays), 1 ml reactions were performed in 1.5 ml

Eppendorf tubes with 50 rpm end-over-end rotation at 25°C. For monitoring oligomer formation, reactions were staggered such that all time points were collected at the same time. Fifty µl of reactions containing 2.5 µM protein were applied to nitrocellulose using a vacuum manifold. Blots were developed using anti-His6 (1:2000 dilution, Invitrogen) or A11 polyclonal antibodies (1:100 dilution) essentially as described [(46)].

To generate amyloid seeds for comparisons of fiber elongation rates, proteins were assembled using continuous end-over-end agitation for 5 days. Aggregates were then collected by ultracentrifugation (100,000 rcf for 1 hour) and resuspended in fresh assembly buffer. We sonicated these preparations to fragment the amyloid fibrils into similar lengths, as determined by SDD-AGE (Figure 4.12A), thus ensuring approximately the same number of fiber ends per mass of polymer. Seed stocks were then normalized according to their amounts of SDS-resistant protein as described [(92)] (Figure 4.12B).

To ensure meaningful comparisons between variants for seeded reactions, we analyzed seeded polymerization rates at varying concentrations of soluble protein. We found no evidence for polymerizable soluble oligomers; all reactions had an approximately first-order dependence on soluble protein concentration (Figure 4.12F-G).

4.12.6 Membrane disruption assay

Toxilight Bioassay kit measures leakage of adenylate kinase from the cells to the extracellular medium due to the loss of cell integrity (damage of plasma membrane). 2 x 10^5 SH-SY5Y cells were seeded in 24-well plates and grown overnight in a 1:1 mixture of DMEM and Ham's F12 and 10% FBS. Fresh or pre-aggregated proteins of Sup35 PrD-M-His7 variants (2.5 µM) were prepared in serum-free medium and applied for 12-

15 hours. Cells were briefly spun at 800 rcf and 30 μl of the medium was carefully removed and used for the toxicity assay as recommended by the manufacturer.

4.12.7 Molecular simulations of polyglutamine and polyasparagine

We simulated one or two polyglutamine (polyQ) and polyasparagine (polyN) molecules N-acetyl-$(Gln)_{30}$-N′-Methylamide and N-acetyl-$(Asn)_{30}$-N′-Methylamide; the chains were modeled in atomic detail; for brevity we refer to these molecules as Q_{30} and N_{30}, respectively. Markov chain Metropolis Monte Carlo (MC) Simulations were performed in the canonical ensemble and molecules were enclosed in a spherical droplet of radius 200Å, which was enforced using a harmonic boundary potential. The replica exchange method was used to enhance conformational sampling [(93)]. The degrees of freedom were the backbone ϕ, ψ, ω and sidechain χ dihedral angles. For MC simulations with two chains, rigid-body coordinates, namely center-of-mass translations and rotations were included as additional degrees of freedom. Bond lengths and bond angles were held fixed at values prescribed by Engh and Huber [(94)]. We used parameters from the OPLS-AA/L forcefield [(95)] with appropriate modifications and the ABSINTH implicit solvent model [(96)]. Details of the move sets, the sampling protocol, and convergence tests are identical to those used in previous work on similar systems [(47)].

Simulations were performed in the presence or absence of local conformational restraints. To impose conformational restraints, we used a parameter referred to as f_β, which denotes the fraction of residues in the polypeptide that are biased to sample backbone dihedral angles from the β-basin of (ϕ,ψ) space; f_β assumes values between 0 and 1. The method used to quantify f_β has been described in published work [(47)]. Local dihedral angle biases were incorporated by adding a harmonic restraint potential of the

form $U_{restr}= k(f_\beta - f_{\beta,0})^2$ to the molecular mechanics energy functions. Here, $f_{\beta,0}$ and f_β are the target and actual values for f_β, respectively. Following calibrations performed in previous work, k=2.5 kcal/mol per restrained degree of freedom for both N_{30} and Q_{30}. When $f_{b,0}$ =1, the backbone ϕ, ψ angles for all residues in the polypeptides are biased to adopt conformations from the β-basin. The entropic penalty associated with sampling ϕ, ψ angles from the β-basin is pre-paid in simulations carried out in the presence of restraints placed on f_b . The extent of formation of ordered β-sheets is quantified using normalized DSSP-E scores [97], which are used to quantify β-sheet contents in protein structures.

4.12.8 Molecular simulations of turn forming regions in Sup35

The following sequence fragments from Sup35 were simulated: QQYNPQGGYQ and QQYNPDAGYQ including mutations of all Q's to N's and all N's to Q's for both sequences. As with the polyasparagine and polyglutamine molecules, these constructs were simulated in atomic detail with the ABSINTH implicit solvent model [96]. No conformational restraints or enhanced sampling techniques were necessary. The moveset and sampled degrees of freedom were the same as discussed previously. Data were accumulated over 10 independent MC simulations performed at T=298 K. Turn dihedral angles were defined as the dihedral angle calculated using the four α-carbon atoms from residues NPQG and NPDA. The distance between α-carbon atoms 1 and 4 were defined as residue pairs N to G for the NPQG motif and N to A for the NPDA motif. The two-dimensional histograms of these quantities had a bin size of 4 degrees and 0.3Å respectively. All instances of the turn motifs NPQG and NPDA were collected from the protein data bank (PDB), current as of February 2011, and the dihedral and distance

quantities were calculated for each occurrence and annotated on top of the two-dimensional histogram. The PROSS secondary structure assignment tool was used to classify each of these annotations as β-turn or coil. These motifs were not a part of any other type of secondary structure element.

4.13 References

1. Halfmann, R., Alberti, S., Krishnan, R., Lyle, N., O'Donnell, C. W., King, O. D., Berger, B., Pappu, R. V., and Lindquist, S. (2011) Opposing Effects of Glutamine and Asparagine Govern Prion Formation by Intrinsically Disordered Proteins, *Molecular Cell 43*, 72-84.
2. Radivojac, P., Iakoucheva, L. M., Oldfield, C. J., Obradovic, Z., Uversky, V. N., and Dunker, A. K. (2007) Intrinsic disorder and functional proteomics, *Biophys J 92*, 1439-1456.
3. Turoverov, K. K., Kuznetsova, I. M., and Uversky, V. N. (2010) The protein kingdom extended: ordered and intrinsically disordered proteins, their folding, supramolecular complex formation, and aggregation, *Prog Biophys Mol Biol 102*, 73-84.
4. Fuxreiter, M., Tompa, P., Simon, I., Uversky, V. N., Hansen, J. C., and Asturias, F. J. (2008) Malleable machines take shape in eukaryotic transcriptional regulation, *Nat Chem Biol 4*, 728-737.
5. Vavouri, T., Semple, J. I., Garcia-Verdugo, R., and Lehner, B. (2009) Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity, *Cell 138*, 198-208.

6. Gsponer, J., Futschik, M. E., Teichmann, S. A., and Babu, M. M. (2008) Tight Regulation of Unstructured Proteins: From Transcript Synthesis to Protein Degradation, *Science 322*, 1365-1368.

7. Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., and Dunker, A. K. (2001) Sequence complexity of disordered protein, *Proteins 42*, 38-48.

8. Weathers, E. A., Paulaitis, M. E., Woolf, T. B., and Hoh, J. H. (2004) Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein, *FEBS Lett 576*, 348-352.

9. Toombs, J. A., McCarty, B. R., and Ross, E. D. (2010) Compositional determinants of prion formation in yeast, *Mol Cell Biol 30*, 319-332.

10. Pierce, M. M., Baxa, U., Steven, A. C., Bax, A., and Wickner, R. B. (2005) Is the prion domain of soluble Ure2p unstructured?, *Biochemistry 44*, 321-328.

11. Alberti, S., Halfmann, R., King, O., Kapila, A., and Lindquist, S. (2009) A systematic survey identifies prions and illuminates sequence features of prionogenic proteins, *Cell 137*, 146-158.

12. Perutz, M. F., Pope, B. J., Owen, D., Wanker, E. E., and Scherzinger, E. (2002) Aggregation of proteins with expanded glutamine and alanine repeats of the glutamine-rich and asparagine-rich domains of Sup35 and of the amyloid beta-peptide of amyloid plaques, *Proc Natl Acad Sci U S A 99*, 5596-5600.

13. Uversky, V. N. (2008) Amyloidogenesis of natively unfolded proteins, *Curr Alzheimer Res 5*, 260-287.

14. Nelson, R., and Eisenberg, D. (2006) Recent atomic models of amyloid fibril structure, *Curr Opin Struct Biol 16*, 260-265.

15. Serio, T. R., Cashikar, A. G., Kowal, A. S., Sawicki, G. J., Moslehi, J. J., Serpell, L., Arnsdorf, M. F., and Lindquist, S. L. (2000) Nucleated conformational conversion and the replication of conformational information by a prion determinant, *Science 289*, 1317-1321.

16. Walters, R. H., and Murphy, R. M. (2009) Examining polyglutamine peptide length: a connection between collapsed conformations and increased aggregation, *J Mol Biol 393*, 978-992.

17. Williamson, T. E., Vitalis, A., Crick, S. L., and Pappu, R. V. (2010) Modulation of polyglutamine conformations and dimer formation by the N-terminus of huntingtin, *J Mol Biol 396*, 1295-1309.

18. Mukhopadhyay, S., Krishnan, R., Lemke, E. A., Lindquist, S., and Deniz, A. A. (2007) A natively unfolded yeast prion monomer adopts an ensemble of collapsed and rapidly fluctuating structures, *Proc Natl Acad Sci U S A 104*, 2649-2654.

19. Krishnan, R., and Lindquist, S. L. (2005) Structural insights into a yeast prion illuminate nucleation and strain diversity, *Nature 435*, 765-772.

20. Perutz, M. F., and Windle, A. H. (2001) Cause of neural death in neurodegenerative diseases attributable to expansion of glutamine repeats, *Nature 412*, 143-144.

21. Dueholm, M. S., Petersen, S. V., Sonderkaer, M., Larsen, P., Christiansen, G., Hein, K. L., Enghild, J. J., Nielsen, J. L., Nielsen, K. L., Nielsen, P. H., and Otzen, D. E. (2010) Functional amyloid in Pseudomonas, *Mol Microbiol*.

22. Larsen, P., Nielsen, J. L., Dueholm, M. S., Wetzel, R., Otzen, D., and Nielsen, P. H. (2007) Amyloid adhesins are abundant in natural biofilms, *Environ Microbiol 9*, 3077-3090.

23. Hammer, N. D., Wang, X., McGuffie, B. A., and Chapman, M. R. (2008) Amyloids: friend or foe?, *J Alzheimers Dis 13*, 407-419.

24. Wang, X., Hammer, N. D., and Chapman, M. R. (2008) The molecular basis of functional bacterial amyloid polymerization and nucleation, *J Biol Chem 283*, 21530-21539.

25. Glover, J. R., Kowal, A. S., Schirmer, E. C., Patino, M. M., Liu, J. J., and Lindquist, S. (1997) Self-seeded fibers formed by Sup35, the protein determinant of [PSI+], a heritable prion-like factor of S. cerevisiae, *Cell 89*, 811-819.

26. Halfmann, R., and Lindquist, S. (2010) Epigenetics in the extreme: prions and the inheritance of environmentally acquired traits, *Science 330*, 629-632.

27. True, H. L., and Lindquist, S. L. (2000) A yeast prion provides a mechanism for genetic variation and phenotypic diversity, *Nature 407*, 477-483.

28. Halfmann, R., Alberti, S., and Lindquist, S. (2010) Prions, protein homeostasis, and phenotypic diversity, *Trends Cell Biol 20*, 125-133.

29. Michelitsch, M. D., and Weissman, J. S. (2000) A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions, *Proc Natl Acad Sci U S A 97*, 11910-11915.

30. Ross, E. D., Edskes, H. K., Terry, M. J., and Wickner, R. B. (2005) Primary sequence independence for prion formation, *Proc Natl Acad Sci U S A 102*, 12825-12830.

31. Osherovich, L. Z., Cox, B. S., Tuite, M. F., and Weissman, J. S. (2004) Dissection and design of yeast prions, *PLoS Biol 2*, E86.

32. Lancaster, A. K., Bardill, J. P., True, H. L., and Masel, J. (2010) The spontaneous appearance rate of the yeast prion [PSI+] and its implications for the evolution of the evolvability properties of the [PSI+] system, *Genetics 184*, 393-400.

33. Derkatch, I. L., Chernoff, Y. O., Kushnirov, V. V., Inge-Vechtomov, S. G., and Liebman, S. W. (1996) Genesis and variability of [PSI] prion factors in Saccharomyces cerevisiae, *Genetics 144*, 1375-1386.

34. Chernoff, Y. O., Lindquist, S. L., Ono, B., Inge-Vechtomov, S. G., and Liebman, S. W. (1995) Role of the chaperone protein Hsp104 in propagation of the yeast prion-like factor [psi+], *Science 268*, 880-884.

35. Grimminger, V., Richter, K., Imhof, A., Buchner, J., and Walter, S. (2004) The prion curing agent guanidinium chloride specifically inhibits ATP hydrolysis by Hsp104, *J Biol Chem 279*, 7378-7383.

36. Kryndushkin, D. S., Alexandrov, I. M., Ter-Avanesyan, M. D., and Kushnirov, V. V. (2003) Yeast [PSI+] prion aggregates are formed by small Sup35 polymers fragmented by Hsp104, *J Biol Chem 278*, 49636-49643.

37. Derkatch, I. L., Bradley, M. E., Hong, J. Y., and Liebman, S. W. (2001) Prions affect the appearance of other prions: the story of [PIN(+)], *Cell 106*, 171-182.

38. LeVine, H., 3rd. (1993) Thioflavine T interaction with synthetic Alzheimer's disease beta-amyloid peptides: detection of amyloid aggregation in solution, *Protein Sci 2*, 404-410.

39. DiFiglia, M., Sapp, E., Chase, K. O., Davies, S. W., Bates, G. P., Vonsattel, J. P., and Aronin, N. (1997) Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neurites in brain, *Science 277*, 1990-1993.

40. Meriin, A. B., Zhang, X., He, X., Newnam, G. P., Chernoff, Y. O., and Sherman, M. Y. (2002) Huntington toxicity in yeast model depends on polyglutamine aggregation mediated by a prion-like protein Rnq1, *J Cell Biol 157*, 997-1004.

41. Krobitsch, S., and Lindquist, S. (2000) Aggregation of huntingtin in yeast varies with the length of the polyglutamine expansion and the expression of chaperone proteins, *Proc Natl Acad Sci U S A 97*, 1589-1594.

42. Duennwald, M. L., Jagadish, S., Giorgini, F., Muchowski, P. J., and Lindquist, S. (2006) A network of protein interactions determines polyglutamine toxicity, *Proc Natl Acad Sci U S A 103*, 11051-11056.

43. Kawai-Noma, S., Pack, C. G., Kojidani, T., Asakawa, H., Hiraoka, Y., Kinjo, M., Haraguchi, T., Taguchi, H., and Hirata, A. (2010) In vivo evidence for the fibrillar structures of Sup35 prions in yeast cells, *J Cell Biol 190*, 223-231.

44. Tyedmers, J., Treusch, S., Dong, J., McCaffery, J. M., Bevis, B., and Lindquist, S. (2010) Prion induction involves an ancient system for the sequestration of aggregated proteins and heritable changes in prion fragmentation, *Proc Natl Acad Sci U S A 107*, 8633-8638.

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

46. Shorter, J., and Lindquist, S. (2004) Hsp104 catalyzes formation and elimination of self-replicating Sup35 prion conformers, *Science 304*, 1793-1797.

47. Vitalis, A., Lyle, N., and Pappu, R. V. (2009) Thermodynamics of beta-sheet formation in polyglutamine, *Biophys J 97*, 303-311.

48. Vitalis, A., Wang, X., and Pappu, R. V. (2008) Atomistic simulations of the effects of polyglutamine chain length and solvent quality on conformational equilibria and spontaneous homodimerization, *J Mol Biol 384*, 279-297.

49. Vitalis, A., Wang, X., and Pappu, R. V. (2007) Quantitative characterization of intrinsic disorder in polyglutamine: insights from analysis based on polymer theories, *Biophys J 93*, 1923-1937.

50. Pappu, R. V., Wang, X., Vitalis, A., and Crick, S. L. (2008) A polymer physics perspective on driving forces and mechanisms for protein aggregation, *Arch Biochem Biophys 469*, 132-141.

51. Wang, X., Vitalis, A., Wyczalkowski, M. A., and Pappu, R. V. (2006) Characterizing the conformational ensemble of monomeric polyglutamine, *Proteins 63*, 297-311.

52. Zhang, J., and Muthukumar, M. (2009) Simulations of nucleation and elongation of amyloid fibrils, *The Journal of Chemical Physics 130*, 035102-035117.

53. Finkelstein, A. V. (1991) Rate of beta-structure formation in polypeptides, *Proteins 9*, 23-27.

54. Srinivasan, R., and Rose, G. D. (1999) A physical basis for protein secondary structure, *Proceedings of the National Academy of Sciences 96*, 14258-14263.

55. Haynes, C., Oldfield, C. J., Ji, F., Klitgord, N., Cusick, M. E., Radivojac, P., Uversky, V. N., Vidal, M., and Iakoucheva, L. M. (2006) Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes, *PLoS Comput Biol 2*, e100.

56. Chen, S., Ferrone, F. A., and Wetzel, R. (2002) Huntington's disease age-of-onset linked to polyglutamine aggregation nucleation, *Proc Natl Acad Sci U S A 99*, 11884-11889.

57. Powers, E. T., and Powers, D. L. (2008) Mechanisms of protein fibril formation: nucleated polymerization with competing off-pathway aggregation, *Biophys J 94*, 379-391.

58. McDonnell, A. V., Jiang, T., Keating, A. E., and Berger, B. (2006) Paircoil2: improved prediction of coiled coils from sequence, *Bioinformatics 22*, 356-358.

59. Lupas, A., Van Dyke, M., and Stock, J. (1991) Predicting coiled coils from protein sequences, *Science 252*, 1162-1164.

60. Jenkins, J., and Pickersgill, R. (2001) The architecture of parallel beta-helices and related folds, *Prog Biophys Mol Biol 77*, 111-175.

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

62. Tessier, P. M., and Lindquist, S. (2009) Unraveling infectious structures, strain variants and species barriers for the yeast prion [PSI+], *Nat Struct Mol Biol 16*, 598-605.

63. Dong, J., Castro, C. E., Boyce, M. C., Lang, M. J., and Lindquist, S. (2010) Optical trapping with high forces reveals unexpected behaviors of prion fibrils, *Nat Struct Mol Biol 17*, 1422-1430.

64. Patel, B. K., Gavin-Smyth, J., and Liebman, S. W. (2009) The yeast global transcriptional co-repressor protein Cyc8 can propagate as a prion, *Nat Cell Biol*.

65. Salazar, A. M., Silverman, E. J., Menon, K. P., and Zinn, K. (2010) Regulation of synaptic Pumilio function by an aggregation-prone domain, *J Neurosci 30*, 515-522.

66. Si, K., Lindquist, S., and Kandel, E. R. (2003) A neuronal isoform of the aplysia CPEB has prion-like properties, *Cell 115*, 879-891.

67. Decker, C. J., Teixeira, D., and Parker, R. (2007) Edc3p and a glutamine/asparagine-rich domain of Lsm4p function in processing body assembly in Saccharomyces cerevisiae, *J Cell Biol 179*, 437-449.

68. Ross, E. D., Minton, A., and Wickner, R. B. (2005) Prion domains: sequences, structures and interactions, *Nat Cell Biol 7*, 1039-1044.

69. Fernandez-Escamilla, A. M., Rousseau, F., Schymkowitz, J., and Serrano, L. (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins, *Nat Biotechnol 22*, 1302-1306.

70. Tartaglia, G. G., and Vendruscolo, M. (2008) The Zyggregator method for predicting protein aggregation propensities, *Chem Soc Rev 37*, 1395-1401.

71. Maji, S. K., Perrin, M. H., Sawaya, M. R., Jessberger, S., Vadodaria, K., Rissman, R. A., Singru, P. S., Nilsson, K. P., Simon, R., Schubert, D., Eisenberg, D., Rivier, J., Sawchenko, P., Vale, W., and Riek, R. (2009) Functional amyloids as

natural storage of peptide hormones in pituitary secretory granules, *Science 325*, 328-332.

72. Fowler, D. M., Koulov, A. V., Balch, W. E., and Kelly, J. W. (2007) Functional amyloid--from bacteria to humans, *Trends Biochem Sci 32*, 217-224.

73. Harrison, P. M., and Gerstein, M. (2003) A method to assess compositional bias in biological sequences and its application to prion-like glutamine/asparagine-rich domains in eukaryotic proteomes, *Genome Biol 4*, R40.

74. Singh, G. P., Chandra, B. R., Bhattacharya, A., Akhouri, R. R., Singh, S. K., and Sharma, A. (2004) Hyper-expansion of asparagines correlates with an abundance of proteins with prion-like domains in Plasmodium falciparum, *Mol Biochem Parasitol 137*, 307-319.

75. He, B., Wang, K., Liu, Y., Xue, B., Uversky, V. N., and Dunker, A. K. (2009) Predicting intrinsic disorder in proteins: an overview, *Cell Res 19*, 929-949.

76. Kreil, D. P., and Kreil, G. (2000) Asparagine repeats are rare in mammalian proteins, *Trends Biochem Sci 25*, 270-271.

77. Karlin, S., Brocchieri, L., Bergman, A., Mrazek, J., and Gentles, A. J. (2002) Amino acid runs in eukaryotic proteomes and disease associations, *Proc Natl Acad Sci U S A 99*, 333-338.

78. Kozlowski, P., de Mezer, M., and Krzyzosiak, W. J. (2010) Trinucleotide repeats in human genome and exome, *Nucleic Acids Res 38*, 4027-4039.

79. Bacolla, A., Larson, J. E., Collins, J. R., Li, J., Milosavljevic, A., Stenson, P. D., Cooper, D. N., and Wells, R. D. (2008) Abundance and length of simple repeats

in vertebrate genomes are determined by their structural properties, *Genome Res 18*, 1545-1553.

80. Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K., and Obradovic, Z. (2006) Length-dependent prediction of protein intrinsic disorder, *BMC Bioinformatics 7*, 208.

81. Liu, J. J., and Lindquist, S. (1999) Oligopeptide-repeat expansions modulate 'protein-only' inheritance in yeast, *Nature 400*, 573-576.

82. Buchan, J. R., Muhlrad, D., and Parker, R. (2008) P bodies promote stress granule assembly in Saccharomyces cerevisiae, *J Cell Biol 183*, 441-455.

83. Meriin, A. B., Zhang, X., Alexandrov, I. M., Salnikova, A. B., Ter-Avanesian, M. D., Chernoff, Y. O., and Sherman, M. Y. (2007) Endocytosis machinery is involved in aggregation of proteins with expanded polyglutamine domains, *FASEB J 21*, 1915-1925.

84. Titz, B., Thomas, S., Rajagopala, S. V., Chiba, T., Ito, T., and Uetz, P. (2006) Transcriptional activators in yeast, *Nucleic Acids Res 34*, 955-967.

85. Xiao, H., and Jeang, K. T. (1998) Glutamine-rich domains activate transcription in yeast Saccharomyces cerevisiae, *J Biol Chem 273*, 22873-22876.

86. Treusch, S., Cyr, D. M., and Lindquist, S. (2009) Amyloid deposits: protection against toxic protein species?, *Cell Cycle 8*, 1668-1674.

87. Truant, R., Atwal, R. S., Desmond, C., Munsie, L., and Tran, T. (2008) Huntington's disease: revisiting the aggregation hypothesis in polyglutamine neurodegenerative diseases, *FEBS J 275*, 4252-4262.

88. Takahashi, T., Kikuchi, S., Katada, S., Nagai, Y., Nishizawa, M., and Onodera, O. (2008) Soluble polyglutamine oligomers formed prior to inclusion body formation are cytotoxic, *Hum Mol Genet 17*, 345-356.

89. Krishnan, R., Goodman, J. L., Mukhopadhyay, S., Pacheco, C. D., Lemke, E. A., Deniz, A. A., and Lindquist, S. (2012) Conserved features of intermediates in amyloid assembly determine their benign or toxic states, *Proceedings of the National Academy of Sciences*.

90. Luheshi, L. M., Tartaglia, G. G., Brorsson, A. C., Pawar, A. P., Watson, I. E., Chiti, F., Vendruscolo, M., Lomas, D. A., Dobson, C. M., and Crowther, D. C. (2007) Systematic in vivo analysis of the intrinsic determinants of amyloid Beta pathogenicity, *PLoS Biol 5*, e290.

91. Duennwald, M. L., Jagadish, S., Muchowski, P. J., and Lindquist, S. (2006) Flanking sequences profoundly alter polyglutamine toxicity in yeast, *Proc Natl Acad Sci U S A 103*, 11045-11050.

92. Dong, J., Bloom, J. D., Goncharov, V., Chattopadhyay, M., Millhauser, G. L., Lynn, D. G., Scheibel, T., and Lindquist, S. (2007) Probing the role of PrP repeats in conformational conversion and amyloid assembly of chimeric yeast prions, *J Biol Chem 282*, 34204-34212.

93. Sugita, Y., and Okamoto, Y. (1999) Replica-exchange molecular dynamics method for protein folding, *Chemical Physics Letters 314*, 141-151.

94. Engh, R. A., and Huber, R. (1991) Accurate Bond and Angle Parameters for X-Ray Protein-Structure Refinement, *Acta Crystallographica Section A 47*, 392-400.

95. Kaminski, G. A., Friesner, R. A., Tirado-Rives, J., and Jorgensen, W. L. (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides, *Journal Of Physical Chemistry B 105*, 6474-6487.

96. Vitalis, A., and Pappu, R. V. (2009) ABSINTH: A New Continuum Solvation Model for Simulations of Polypeptides in Aqueous Solutions, *Journal of Computational Chemistry 30*, 673-699.

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

Chapter 5

Consequences of N-terminal flanking residues on intra and intermolecular interactions of polyglutamine and their defining role in polyglutamine aggregation mechanisms

This chapter is a culmination of effort spanning several individuals from the Pappu lab. Scott Crick, a former graduate student, provided the anisotropy data and was extremely helpful during discussions regarding these results. Nil Gural, an undergraduate assistant to Nicholas Lyle, the candidate, performed the simulations of homodimers with two disease related flanking sequences (ataxins) and assisted in the associativity and energy decomposition analysis. Simulation results containing the N-terminal flanking sequence from huntingtin were taken from previous results generated by Tim Williamson, a former graduate student. Nicholas Lyle performed the remaining simulations, analysis, and interpretation.

5.1 Introduction

Expanded polyglutamine tracts within specific proteins are associated with at least nine neurodegenerative disorders [(1)] including Huntington's Disease (HD), Spinal Bulbar Muscular Atrophy (SBMA), Spinocerebellar Ataxia (SCA), Dentatorubral-Pallidoluysian Atrophy (DRPLA). All these diseases result in the accumulation of polyglutamine-rich protein aggregates at distinct intracellular locations. Proteins with polyglutamine expansions are destabilized and the length of the polyglutamine expansion determines the extent of destabilization [(2)]. This makes proteins with polyglutamine expansions susceptible to proteolysis [(3-6)]. The proteolytic fragments are rich in polyglutamine and are prone to form aggregates. Indeed, large fractions of the mutant protein's sequence are absent in polyglutamine-rich neuronal intranuclear inclusions [(7)].

Suppression of proteolytic activity in cell models served to increase cell viability [8]. Products of proteolysis are likely to be rich in polyglutamine; it is also likely that they contain some fraction of the N- and C-terminal flanking from the host protein. Cleavage products are difficult to characterize experimentally and hence the precise identity of these fragments is unknown, although it is reasonable to expect some heterogeneity in sequence composition – a heterogeneity that is dependent on the polyglutamine length [9, 10].

Figure 5.5.1 shows evidence of an inverse relationship between the lengths of polyglutamine expansion and ages of onset in different polyglutamine expansion disorders [11]. Given the enriched polyglutamine content of the supposed toxic proteolytic fragments and no relation between the expanded host proteins, it is clear that the intrinsic properties of polyglutamine are involved in disease. However, Figure 5.5.1 makes the point that the threshold age of onset is altered by the identity of the host protein. This is also consistent with several *in vivo* studies that show differential pathogenic effects in animal or cellular models of polyglutamine repeat diseases upon the expression of different sequence constructs [6, 8, 12-15]. These differences are most likely tied to the identity of the sequences flanking the polyglutamine expansion. This conjecture is consistent with experiments probing specific aspects of aggregation kinetics of proteins with polyglutamine expansions and suggests that focusing entirely on the biophysics of homopolymeric polyglutamine is unlikely to yield a complete understanding of the relationship between the aggregation of polyglutamine-rich peptides and disease.

Several experimental studies demonstrate that soluble, monomeric polyglutamine-rich constructs have spectroscopic signals that indicate a lack of well-defined secondary structures and this holds true irrespective of polyglutamine length [16-19]. However, the polyglutamine-rich insoluble neuronal nuclear inclusions associated with polyglutamine expansion diseases are

fibrillar, amyloid-like aggregates that are proposed to have high β-sheet contents [20, 21]. Fiber diffraction data are consistent with the view that individual polyglutamine molecules adopt flat β-sheet architectures in these aggregates [22]. *In vitro* studies have shown that the overall rate of formation of these ordered aggregates increase with polyglutamine length [16, 23, 24]. Chen et al. proposed a model to connect intrinsically disordered monomers and β-sheet structures that are prominent in fibrillar, insoluble aggregates[23]. Analysis of kinetic data on the aggregation of $K_2Q_NK_2$ peptide constructs using a homogeneous nucleation model yielded a nucleus size of one. They interpreted these results as follows: monomeric polyglutamine is in pre-equilibrium between a folded, toxic, β-sheet structure and the disordered ensemble. Whenever the toxic fold is adopted, the thermodynamically unfavorable nucleus is populated and the resulting ordered conformation is elongated through monomer addition. This model, referred to as homogeneous nucleation, proposes that the length dependence of polyglutamine aggregation derives from the increasing stability of the folded monomer nucleus with chain length.

Chapters 3 and 4 provide an alternative explanation for the chain length dependence of aggregation and shows this need not be a nucleated process. For homopolymers, there is a driving force for aggregation if polymers are in milieus that are so-called poor solvents [18]. In such environments, polymers form homogeneously mixed solutions of isolated globules under dilute solution conditions. In the single molecule limit, intrachain interactions are preferred to chain-solvent interactions for chains in a poor solvent, and chain sizes measured using radii of gyration (R_g) or hydrodynamic radii (R_h) scale as $N^{1/3}$ with chain length (N) [25-27]. As a result, the conformational ensemble in dilute solutions is composed of compact, roughly spherical conformations that minimize the interface between chain molecules and the surrounding solvent. In poor solvents, the stabilities of collapsed structures and the driving force for homotypic

intermolecular associations increases with N [28-35]. Crick et al. used fluorescence correlation spectroscopy to show that R_h for monomeric polyglutamine in aqueous solutions scales as $N^{1/3}$ with N, thereby establishing that water is a poor solvent for polyglutamine [36]. Atomistic simulations have yielded insights regarding the conformational equilibria and monomer-dimer equilibria for homopolymeric polyglutamine as a function of chain length and temperature [17, 37]. The main findings are as follows: 1) Monomeric polyglutamine forms collapsed, disordered ensembles independent of chain length. 2) The stability of globular forms increases with chain length. 3) The stability of homodimers is coupled to the stability of monomeric globules. 4) Homodimerization – the incipient step in aggregation – does not require specific, nucleated conformational changes. 5) Spontaneous fluctuations, specifically the liquid-like nature of the surfaces of polyglutamine globules are important for promoting intermolecular associations whereby collapsed polyglutamine molecules further reduce their interfaces with the surrounding poor solvent by forming aggregates.

As illustrated in Figure 5.1, the identity of the mutant protein containing the polyglutamine insertion alters the age of onset of different diseases [11]. Since proteolytic fragments are most likely to be involved in disease pathology [38], the peptide sequences immediately flanking the polyglutamine expansion must modulate intrinsic polyglutamine conformational preferences and intermolecular associations [14]. Studies from Wetzel's group show that the C-terminal, 11-residue polyproline segment from huntingtin slows aggregation kinetics vis-à-vis polyglutamine segments sans the proline segment [39]. The origin of this effect is poorly understood. Conversely, similar studies from the Wetzel lab have probed the effects of the Nt17 fragment from huntingtin. The rate of loss of soluble material is enhanced in constructs that have the Nt17 fragment, thereby leading to the proposal that this segment initiates

aggregation [40]. The hypothesis is that interactions between Nt17 lead to an increased effective concentration of polyglutamine segments around each other, thereby increasing the rate of polyglutamine aggregation. Frydman and coworkers reached similar conclusions by quantifying the effects of the Nt17 segment on the rate of accumulation of large linear aggregates that can be captured by an appropriately sized filter trap [41]. Wetzel's results are confounded by the use of polyglutamine constructs with N- and C-terminal lysine residues as reference states for interpreting their data [42]. Frydman's experiments do not probe the presence of spherical, amorphous aggregates that appear to be the norm with homopolymeric molecules [43, 44]. Simulations of Ac-Nt17-Q_N-Nme performed by Williamson et al. demonstrate that the Nt17 segment causes a diminution of intrinsic polyglutamine associations and intrinsic conformational preferences [45]. These results are consistent with Serrano's hypothesis that naturally occurring flanking sequences can act as *gatekeepers* by suppressing intrinsic aggregation propensities of aggregation-prone regions. This chapter expands on these initial studies and quantifies the gatekeeping potentials of different disease-relevant N-terminal sequence contexts. Through comparative studies, we find flanking sequences act as gatekeepers by several mechanisms and these mechanisms are determined by flanking sequence properties. This work allows for the development of a mechanistic framework based on viscoelastic spinodal decomposition that can aid in the development of new inhibitors of polyglutamine-mediated aggregation. To resolve the discrepancies between low-resolution experiments that probe one set of quantities and higher resolution simulations that probe interactions in smaller species, it is important that the two modes of investigation converge toward each other in terms of quantities to assay and species to interrogate. Additionally, it is essential to probe the effects of other flanking sequences to

develop testable predictions regarding the effects of sequence context on polyglutamine conformations and aggregation.

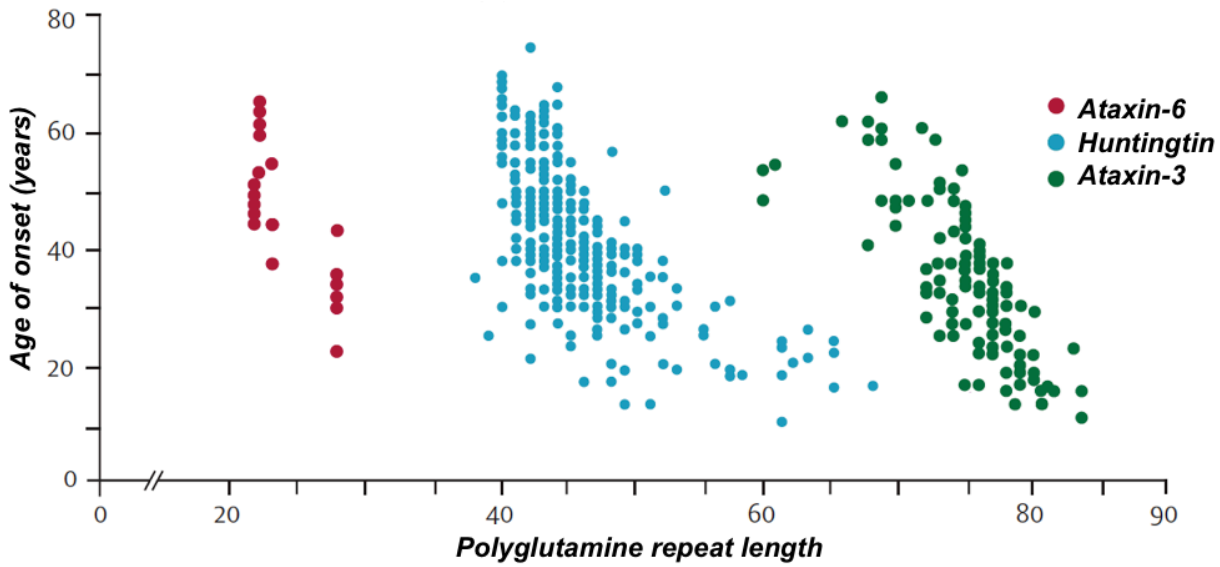


Figure 5.1: Patient data displaying the age of onsets for three human polyglutamine disorders versus polyglutamine expansion length. Data were derived from Figure 3 of Walker et al. All polyglutamine expansion disorders display an inverse correlation between age of onset and expansion length. The age of onset also depends on the polyglutamine flanking sequence that differs in each expansion disease. Spinocerebellar ataxia types 3 and 6 are labeled as Ataxin-3 and Ataxin-6.

This chapter begins with the necessary tests for how well our computational model describes known biophysical properties of synthetic polyglutamine constructs over a range of length scales. Biophysical characterization of polyglutamine peptides has been used to probe the intrinsic properties of polyglutamine and to study the mechanism by which polyglutamine aggregates. These studies use peptide constructs of the form K_2-Q_N-K_2 as proxies for homopolymeric polyglutamine as lysine residues are used to facilitate peptide purification and

solubility [16, 46-49]. Different groups have substituted the flanking lysines for other charged residues [50] or left them off one of the termini [51]. Alterations that modulate solubility might also influence aggregation mechanisms with respect to homopolymeric polyglutamine. In addition to tests of our computational model, we use these results to systematically assess the modulation of aggregation mechanisms, biophysical, and structural properties of polyglutamine by commonly used flanking charges.

5.2 Short synthetic polyglutamine constructs can have rod-like character due to significant sequence dependent α-helical propensities

Polyglutamine tracts form collapsed structures in aqueous solutions to minimize solute-solvent interfaces. For small *N*, α-helical motifs provide access to locally compact structures, and therefore we focused on quantifying the differences between homopolymeric and sequence context-dependent α-helical propensities for polyglutamine. Figure 5.2 shows the ensemble averaged *N*-dependent α-helical contents for polyglutamine tracts in six different sequence constructs including the homopolymer. Table 5.1 summarizes the numerical values for average helical contents and standard deviations calculated over polyglutamine tracts within different sequence constructs. In all constructs, the *N*-dependent α-helical propensities decrease with increasing *N*. It is worth emphasizing that observed helical contents for N=4 convolve the differences in flanking residue identities and contributions due to differences in overall chain length.

For the homopolymeric construct, the average fractional helical content is highest for N=10, and then decreases systematically as the chain becomes long enough to form non-

specifically collapsed structures. While helical propensities become negligibly small with increasing *N* for the homopolymer, there is residual fractional helicity (~0.1) for polyglutamine in constructs with flanking charges. Additionally, the systematic *N*-dependent decreases are different for each of the constructs. For the construct Ac-$K_2CQ_NWK_2$-Nme, the average fractional helical content over the polyglutamine tract decreases from 0.65 (*N*=4) to 0.09 (*N*=24). Alpha helices can be stabilized locally by the presence of one or more appropriately located acidic and basic residues in the N- and C-terminus, respectively. The converse arrangement namely one or more appropriately located acidic and basic residues in the C- and N-terminus, respectively, will have a destabilizing effect on helix nucleation / propagation. The contribution of these "charge-capping" effects was quantified by comparing the average *N*-dependent fractional helical contents for polyglutamine in Ac-$K_2CQ_NWK_2$-Nme to those obtained for polyglutamine in Ac-$D_2Q_NK_2$-Nme and Ac-$G_2Q_NCK_2$-Nme, which are two other constructs used in the experimental literature. Indeed, we find higher α-helical propensities for polyglutamine in Ac-$D_2Q_NK_2$-Nme and Ac-$G_2Q_NCK_2$-Nme (for $N \geq 7$) versus those in Ac-$K_2CQ_NWK_2$-Nme. The analysis of helical contents is relevant in light of the interpretations provided by Singh and Lapidus of their triplet quenching data in constructs of the form $K_2CQ_NWK_2$. They proposed that the polyglutamine segments have rod-like character. Inasmuch as α-helices might be viewed as hydrogen-bonded rods, our results appear to be in partial agreement with the proposals of Singh and Lapidus. However, this rod-like behavior diminishes as *N* increases – a behavior that is in line with expectations from analysis of FRET data reported by Walters and Murphy. A common premise is that the distance between N- and C-terminal lysine residues (for $N \geq 4$) would always be beyond the Bjerrum length. While this might seem like a reasonable assumption, the local, structure making and structure breaking effects of flanking lysine residues must not be

overlooked when analyzing experimental results, especially since the magnitude of these effects varies with N. The results presented in Figure 5.1 make the point that flanking residues used to make polyglutamine tractable for experimental studies can alter the intrinsic conformational preferences vis-à-vis homopolymeric constructs. The extent of alteration decreases with N and short chains within some constructs can show rod-like character due to significant α-helical propensities. Our analysis suggests the need for exercising caution when including results from short chain lengths ($N < 16$) in global analysis of experimental results. Additional caution is needed when extrapolating the findings based on short chain lengths ($N \leq 13$) to length scales relevant to disease ($N \geq 35$). This point is particularly relevant to the hypothesis put forth by Fiumara et al. [52] that coiled-coil interactions are a driver of polyglutamine rich tracts. In the context of our results, the presence of long helical rods is unlikely due to diminishing helical content with an increase in chain length.

It is possible that the alteration to the polyglutamine ensemble due to charged termini could be manifested via non-helical configurations. We find this is not the case because these ensembles are devoid of non-helical elements of secondary structure. For N>16, polymeric properties such as the average radius of gyration $\langle R_g \rangle$ and the end-to-end distance $\langle R_{ee} \rangle$ are consistent with those of a collapsed globule. This follows work from Crick et al. where fluorescence correlation spectrophotometry was used to show synthetic polyglutamine molecules Ac-G$Q_N$$K_2$-Nme are collapsed for N=15,20,24,27,33,36,40,47, and 53.

In summary, the polyglutamine conformational ensemble is heavily influenced the identity of the flanking residues. Because these residues are commonly charged and placed at the termini, helix-capping effects either stabilize or destabilize helicity over the polyglutamine stretch. The magnitude of these alterations are chain length dependent, and their effect

diminishes for N>16. Recent work from Walters and Murphy, described in more detail in the following section, shows that collapse dominates only in the longer length regime N>16, which is what we see here.

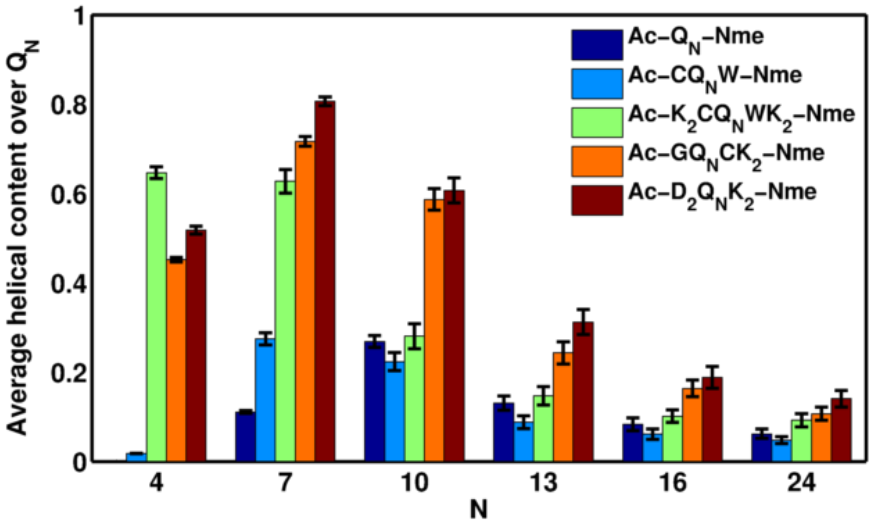


Figure 5.2: Length-dependence of average helical contents calculated over polyglutamine segments in different sequence constructs. Helical content is defined as the fraction of glutamine residues that are part of a π, α, or 310-helix as calculated by the DSSP algorithm. The numerical values for these averages and variances about these values are shown in Table 5.1. Standard errors are shown as error bars. Constructs with N- or C- terminal charges correlate with a higher degree of helical content of which decreases with polyglutamine length (N).

	Q_N*		CQ_NW*		GQ_NCK_2*		$K_2CQ_NWK_2$*		DQ_NK_2*	
N	**⟨H⟩**	**Var(H)**	**⟨H⟩**	**Var(H)**	**⟨H⟩**	**Var(H)**	**⟨H⟩**	**Var(H)**	**⟨H⟩**	**Var(H)**
4	0.00	0.00	0.02	0.00	0.45	0.00	0.65	0.02	0.52	0.01
7	0.11	0.00	0.27	0.02	0.72	0.01	0.63	0.08	0.81	0.01
10	0.27	0.02	0.23	0.05	0.58	0.08	0.28	0.09	0.61	0.11
13	0.13	0.03	0.09	0.02	0.24	0.08	0.15	0.06	0.31	0.11
16	0.08	0.02	0.06	0.02	0.17	0.04	0.10	0.02	0.19	0.08
24	0.06	0.01	0.05	0.01	0.11	0.03	0.09	0.03	0.14	0.04
RMSD†	**0.00**		**0.03**		**0.14**		**0.14**		**0.16**	

Table 5.1: Average helical contents and variances calculated over polyglutamine tracts in different constructs. *All sequences are capped using N-acetyl and N′-methylamide groups

†The root mean squared deviation (RMSD) is calculated with respect to the homopolymeric construct using: $\mathrm{RMSD} = \frac{1}{6}\sqrt{\sum_i \left(\langle H \rangle_i^{\mathrm{X}} - \langle H \rangle_i^{\mathrm{Q}_N} \right)^2}$ where the summation *i* runs over polyglutamine lengths shown in column 1, $\langle H \rangle_i^{\mathrm{X}}$ denotes the average helical content for polyglutamine of length *i* within construct X and $\langle H \rangle_i^{\mathrm{Q}_N}$ denotes the corresponding helical content within the homopolymeric construct of length *i*.

5.3 Comparison of experimentally and computationally derived scaling of polyglutamine end-to-end distances (R_{ee}) with chain length

Figure 5.3 compares FRET derived estimates of $\langle R_{ee} \rangle$ of polyglutamine with charged termini to computational estimates. The sequence constructs Ac-$K_2CQ_NWK_2$-Nme where *N*=7,10,13,16,24 was used to generate the computational results. FRET data from Murphy was available for Ac-$K_2WQ_NXK_2$-Nme constructs with *N*=8,12,16,20,24, where X is either an alanine or dansylated lysine. The data are expressed in terms of the relative change in the length of the

polyglutamine segment, whereby all values are referenced to N=24. Exact numerical agreement between the computational and FRET derived $\langle R_{ee} \rangle$ is not possible due to different reference points used to calculate the length of the polyglutamine segment, i.e., FRET is based on the cystine to dansylated lysine distance where the computational estimate is derived from the cystine to tryptophan distance. The data show that the computationally derived rate at which $\langle R_{ee} \rangle$ increases with N is consistent with the FRET data. The data in Figures 2 and 3 make the point that our computational predictions for the behavior of synthetic polyglutamine constructs are consistent with available biophysical measurements. Importantly, results from our computations are 1) applicable to all length scales explored by synthetic constructs and 2) explain the observation of incomplete collapse for short N<16 synthetic constructs.

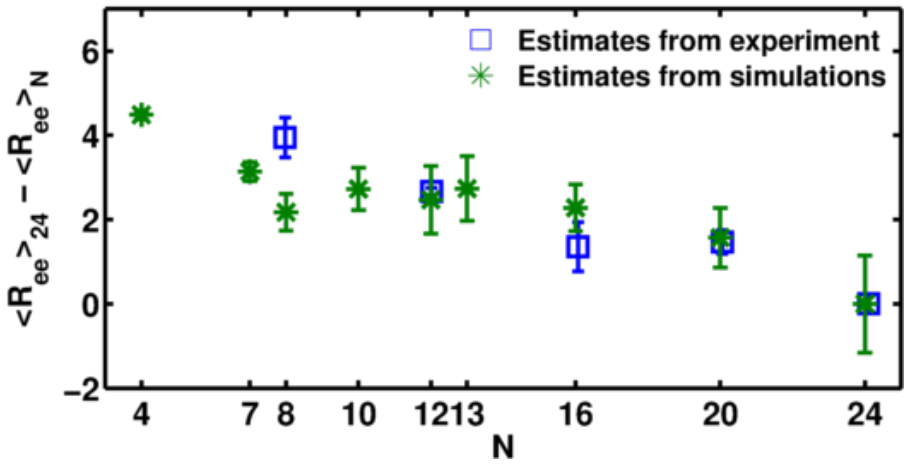


Figure 5.3: Quantitative comparisons of average end-to-end distances. For these comparisons, we used simulation results based on the ABSINTH model for the average distance between cysteine and tryptophan residues in constructs of the form Ac-$K_2CQ_NWK_2$-nme (N=7,10,13,16, and 24). This residue-to-residue distance was defined as the average pairwise distance between all unique inter-residue atom pairs. These were compared to distance estimates obtained by Walters and Murphy from FRET measurements on constructs of the form Ac-K_2WQ_N(dK)K_2-nme (N=8,12,16, 20, and 24), where (dK) denotes a dansylated lysine. The

published FRET data provide estimates of distances between tryptophan and the dansyl group. The plots were designed to provide a visual assessment of how $\langle R_{ee} \rangle$ changes with polyglutamine length, N. We placed results from simulation results and those from experimental measurements on a similar footing by plotting $\langle R_{ee} \rangle_{24} - \langle R_{ee} \rangle_N$ against N, where $\langle R_{ee} \rangle_{24}$ denotes the value for $\langle R_{ee} \rangle$ for N=24. Comparison of the two sets of points demonstrate that the length dependence of $\langle R_{ee} \rangle_{24} - \langle R_{ee} \rangle_N$ estimated using the simulations agree quantitatively with estimates from the FRET measurements.

5.4 Flanking charged termini alter polyglutamine homotypic interactions and conformational heterogeneity

Figure 5.4 shows the 2D probability distribution of asphericity and R collected from molecular simulations of two polyglutamine molecules of length N=35 with and without charged termini. R measures the center-of-mass to center-of-mass distance between the two molecules and asphericity quantifies the departure of a molecule from being spherically symmetric. A value of zero corresponds to a perfect sphere. The synthetic constructs employed were Ac-$Q_{35}K_2$-Nme and Ac-$K_2Q_{35}K_2$-Nme as these are the same constructs used in recent biophysical characterizations of polyglutamine aggregation. Each panel corresponds to a different sequence construct and each region demarcates different types of conformations: Region 1 encompasses the set of all dissociated conformations, i.e., it consists of monomers of any shape; Region 2 consists of dimers that are roughly spherical; and Region 3 contains dimers that have become elongated.

The presence of terminal charges reduces homotypic interactions vis-à-vis the homopolymer and this reduction is dependent on the magnitude of molecular net charge. The integrated probability of Region 1 versus the sum of the integrated probability from Region 2

and Region 3 describes the monomer-dimer equilibrium for each construct. Neutral homopolymers are the least soluble and the total of Region 1 is 25%. Adding two C-terminal lysines increase the probability to 48% and intermolecular associations are almost absent for the quadruply lysinated molecules as the probability increases to 97%. Together these results show charged termini are gatekeepers because they reduce polyglutamine intermolecular interactions and this occurs in the length regime relevant to disease. The gatekeeping mechanism is a consequence of: 1) long-range charge-charge repulsions that increase the barrier for molecular association and 2) terminal residues that eliminate modes of association by preventing burial of charged residues. This reduces the heterogeneity in the types of intermolecular interactions that are possible and leads to preferences in oligomer topologies.

Asphericity quantifies the shape of each molecule and increased variance of this quantity tracks with increased conformational heterogeneity. Figure 5.4 shows that neutral and doubly lysinated constructs in Figure 5.4 can adopt all values of asphericity from 0 to 0.4 with a small fraction (1-3%) outside of this range. This conformational diversity is reduced for the quadruply lysinated construct as values of asphericity generally range from 0 to 0.2. Although most conformations remain spherically symmetric (asphericity < 0.4), flanking charges play a role in the magnitude of fluctuations about spherical symmetry.

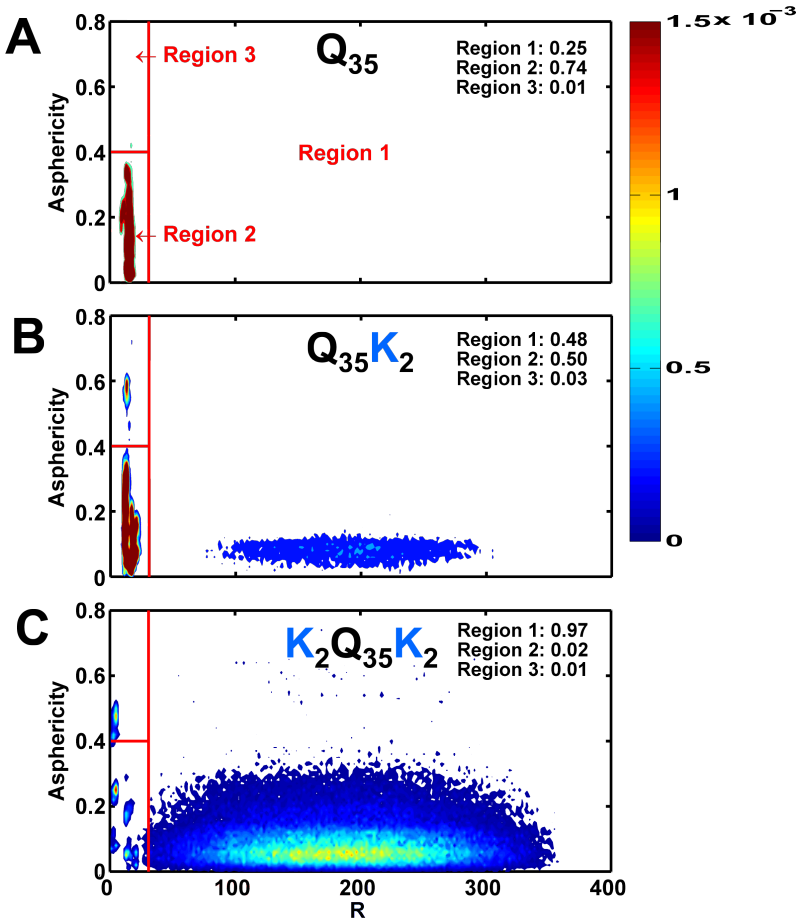


Figure 5.4: Two-dimensional probability distributions of R and asphericity *P*(*R*, asphericity) for synthetic polyglutamine constructs at *T*=315K. *R* denotes the intermolecular distance of separation calculated as the distance between the centers-of-mass for a pair of molecules. Asphericity was calculated for each molecule separately using the eigenvalues of gyration tensors as described in the (methods section). Bin widths of 2Å were used for *R* and 0.02 for asphericity. Legends show occupation statistics for three different intervals. These are: Region 1, [$R \geq 30$Å, Asphericity ≥ 0], Region 2, [$R < 30$Å, Asphericity < 0.4], and Region 3, [$R < 30$Å, Asphericity ≥ 0.4]. Values of Asphericity < 0.4 correspond to spherically shaped globules where values above this undergo significant deformations resulting in either oblate or prolate conformations. $R < 30$Å corresponds to conformations with significant homotypic interactions. Increasing net charge results in preferential occupation of Region 1 consisting of spherical

monomers and a lack of homotypic interactions. Constructs containing no charged residues form roughly spherical dimers (Region 2) with increased fluctuations in Asphericity.

5.5 Flanking charges alter the rate of aggregation and preferred oligomer sizes of synthetic polyglutamine constructs

Figure 5.5 summarizes results from fluorescence anisotropy experiments that compare the effect of changing the pH in a 50 mM chloride free phosphate buffer on the rate of increase of anisotropy of 50 μM synthetic polyglutamine constructs Ac-$K_2Q_{40}K_2$-Nme and Ac-$Q_{40}K_2$-Nme. Anisotropy magnitudes quantify the relative size of a molecule or growing aggregate. Larger anisotropy values correspond to larger aggregate sizes. Increasing the pH of the phosphate buffer from 5 to 9 shifts the primary anionic species from $H_2PO_4^-$ at pH 5 to HPO_4^{2-} at pH 9. At pH 7, the ratio is approximately 1:1. The rate of increase in anisotropy is higher for the Ac-$Q_{40}K_2$-Nme constructs when compared to the corresponding rates for Ac-$K_2Q_{40}K_2$-Nme indicating that higher molecular charge density decreases the rate of aggregation and lengthens the lag time. Additionally, the Ac-$Q_{40}K_2$-Nme peptide with two lysine residues is less sensitive to the buffer pH than the Ac-$K_2Q_{40}K_2$-Nme peptide. The plateau value of the anisotropy is higher for Ac-$Q_{40}K_2$-Nme than for Ac-$K_2Q_{40}K_2$-Nme irrespective of the buffer pH, suggesting that lower molecular charge density allows for soluble oligomers of higher aggregate numbers than those with higher charge density. The dependence of the plateau value on pH, i.e. the sizes of species that remain soluble, is different for the two constructs. Because overall rates for initial cluster formation and preferred oligomeric cluster sizes can be modulated through electrostatic repulsions it is reasonable to expect the presence of flanking lysine residues can alter the aggregation mechanism of polyglutamine-rich constructs.

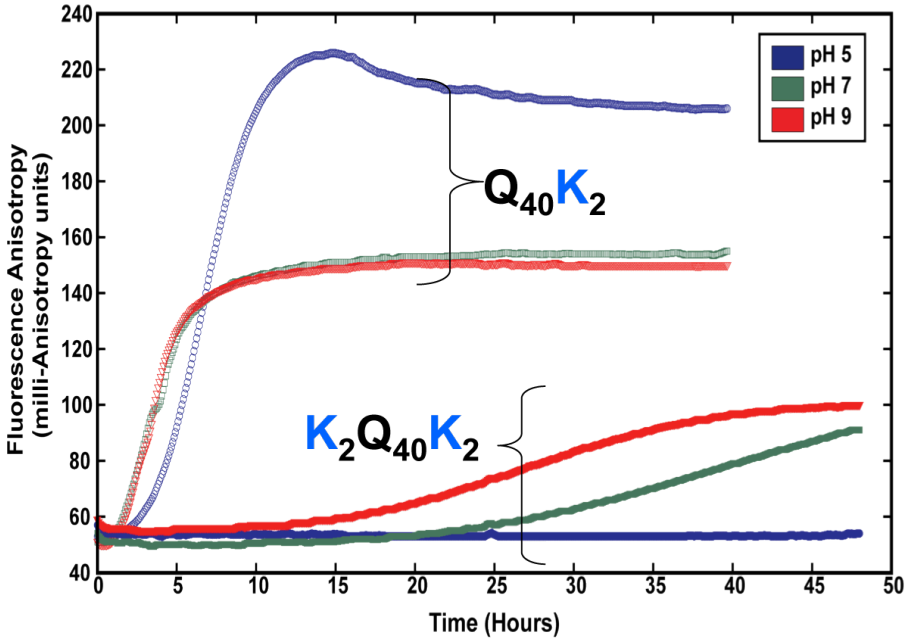


Figure 5.5: Effects of flanking lysine residues on the rate of aggregation as measured using fluorescence anisotropy. Comparative rates of aggregation for two constructs Q40K2 (open symbols) and K2-Q40-K2 (closed symbols) (pH 5, blue), in phosphate buffers of three different pH values (blue, pH 5), (green, pH 7), (red, pH 9).

5.6 N-terminal flanking sequences act as gatekeepers with sequence dependent gatekeeping efficiencies

The preceding results show that synthetic polyglutamine constructs with flanking lysine residues alter the kinetics and thermodynamics of polyglutamine aggregation by reducing intermolecular contacts. Since a difference of only two residues on each termini is sufficient for a gatekeeping effect, it is reasonable that natural flanking sequences found in expansion diseases are capable of modulating polyglutamine aggregation. In fact, ages of onset for polyglutamine expansion diseases cannot be explained solely by expansion length, a point that is evident in Figure 5.1. Aggregation prone products of proteolysis from expanded proteins contain flanking residues; therefore a simple hypothesis is that these flanking residues play a role in modulating

the ages of onsets for these diseases. We propose that part of this modulation is due to changes in intermolecular interactions between expanded proteins and this hypothesis is the focus of this section. We use atomistic simulations to perform comparative assessments of the effects of N-terminal flanking sequences from huntingtin (htt), spinocerebellar ataxia types 3 (atxn-3), and type 6 (atxn-6) because these sequences have distinctive impacts on ages-of-onset. Studies to probe htt C-terminal flanking effects are ongoing and under the direction of Kiersten Ruff, a graduate student in the Pappu lab. N- and C- terminal flanking effects for other expansion diseases are planed for future study.

Sequences for human versions of htt, atxn-3, and atxn-6 were obtained from the UniProtKB database. We used the disorder predictor DRIP-PRED [(53)] to identify the intrinsically disordered regions that encompass polyglutamine. These disordered regions were submitted to the ExPasy proteomics server to identify the most likely proteolytic products of the three sequences by cytoplasmic endopeptidases. The resulting sequences are summarized in Table 5.2 and were the sequence constructs used in the molecular simulations.

Protein Name	*N*-th	Sequence
htt	~35	MA*T*L**EK**LM**K**AF**E***S*L**K***S*F
atxn-3	~80	**EE**L**RKRR**FA*Y*FF**K***QQQ***K**
atxn-6	~23	*H*V*SYS*PVI**RK**A*GGSG*PP

Table 5.2: Sequences of N-terminal segments predicted to be products of proteolysis. Predicted proteolytic products include the polyglutamine stretch between the N- and C-terminal fragments. The second column shows the average threshold lengths (*N*-th) of polyglutamine tracts within each of the three disease-related proteins; the N-terminal segments are all of length 17 derived from the host protein. Single letter notations are used for amino acid residues and the annotation is as follows: polar residues are italicized, residues with charged sidechains at neutral

pH are shown in bold face, hydrophobic residues are underlined and prolyl residues are shown in normal font.

Tim Williamson, a graduate student in the Pappu lab, began work on flanking sequence effects by studying the 17-residue N-terminal htt fragment. This work was later published [54] and revealed that this fragment goes through an order-to-disorder transition as the length of the polyglutamine expansion C-terminal to the fragment was increased. This transition resulted in sequestration and burial of exposed polyglutamine surfaces and this in turn lead to a diminution of the driving force for homotypic interactions. Here, we widen our scope to other flanking sequences to answer the following questions: 1) Do other flanking sequences act as gatekeepers or are some anti-gatekeeping (aggregation enhancing) sequences? 2) What are the mechanisms by which gatekeepers reduce intermolecular associations? 3) Are there rules based on flanking sequence properties that determine gatekeeper efficiencies and mechanisms?

To quantify the gatekeeping efficiency of each flanking sequence, we measured the relative probability of dissociation (P_{rel}). This quantity was derived from molecular simulations of two molecules with the sequence Ac-X17-Q_{35}-Nme and measures the factor by which the N-terminal flanking sequence X17 modulated homotypic interactions with respect to the homopolymer Ac-Q_{35}-Nme. Figure 5.6 shows values of P_{rel} that are either close to or larger than 1, indicating the natural N-terminal flanking sequences are all gatekeepers, but some are more effective than others. From the least to most effective gatekeeping, we find atxn-6 < htt < atxn-3 and this trend correlates with age of onsets with atxn-6 having the earliest and atxn-3 the latest age of onset on average. As shown by Figure 5.6, this trend is robust across the range of temperatures at which intermolecular associations are preferred, i.e., poor solvent conditions.

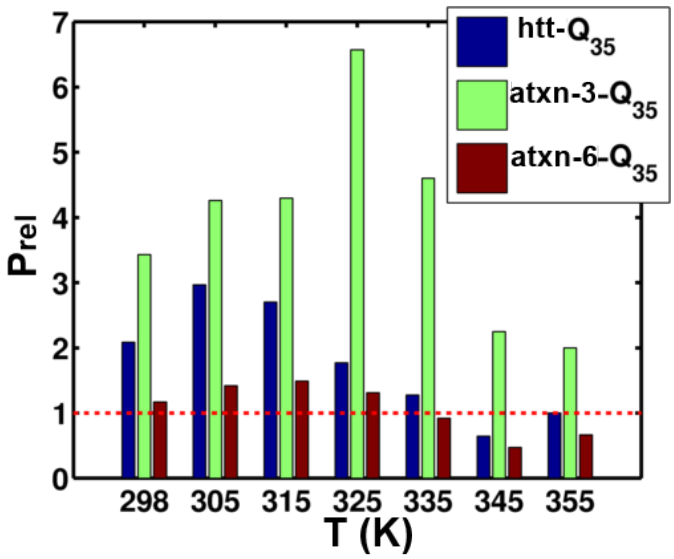


Figure 5.6: Naturally occurring N-terminal flanking sequences have different gatekeeping efficiencies. Two identical molecules of Ac-X17-Q_{35}-Nme were simulated in a droplet of radius of 200Å over a range of temperatures where intermolecular interactions were observed. X17 refers to the naturally occurring 17-residue N-terminal flanking sequence and these were derived from the huntingtin protein (htt), spinocerebellar ataxia types 3 (atxn-3) and type 6 (atxn-6). The full sequences are listed in Table 5.2. The probability of association (P_A) between pairs of molecules is defined as the number of observations of R < 25Å divided by the total number of observations taken over the simulation. R is the pairwise distance between two molecular centers of mass. The relative dissociativity (P_{rel}) is calculated by referencing P_A to P_A for the homopolymer Q_{35}: $P_A = \frac{P_A\ Homopolymer}{P_A\ X17}$. Values for $P_{rel} > 1$ indicate the flanking sequence reduces aggregation with respect to the homopolymer.

The same trends are evident in Figure 5.7 as effective gatekeepers shift probability density from regions 2 and 3 to region 1. As is the case in both Figures 4 and 7, intermolecular interactions correlate with increased values for asphericity and flanking sequences enhance this

effect. Thus, flanking sequences elicit an effect by simultaneous modulation of preferred conformations adopted by the polyglutamine segment and modulation of intermolecular interactions. The degree to which these modulations occur depends on the length of the polyglutamine expansion *N* and the sequence properties of the flanking sequence.

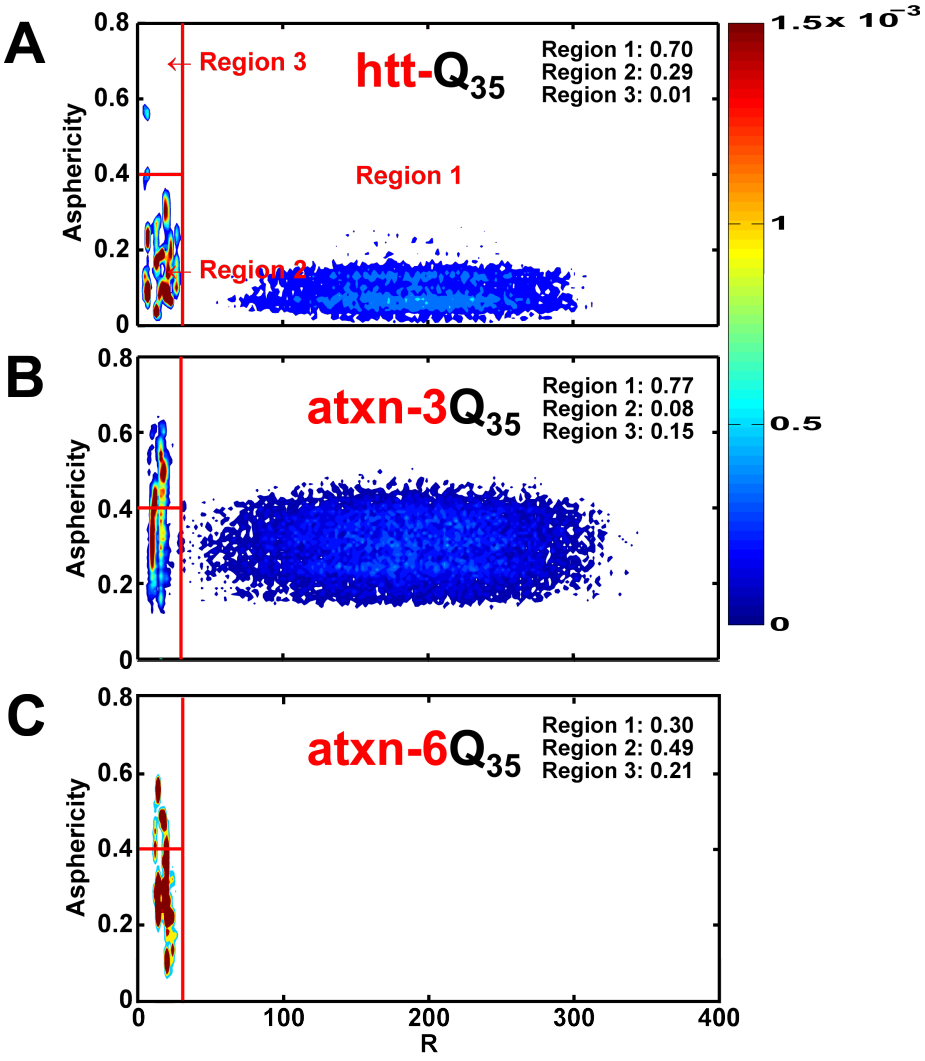


Figure 5.7: *P*(*R*, asphericity) for polyglutamine constructs with N-terminal flanking sequences at *T*=315K. The definition of regions and the construction of the 2D distributions are analogous to that in Figure 5.4. The integrated probability density in region 1 indicates atxn-3 is most effective in preventing intermolecular associations, followed by htt and lastly atxn-6.

Asphericity values indicate atxn-3 and atxn-6 form elongated prolate ellipsoids in addition to spherically symmetric conformations. For all constructs, intermolecular interactions along with the identity of the flanking sequence promote the sampling of elongated conformations.

5.7 Energetics of polyglutamine sequestration

Collapse of homopolymeric polyglutamine increases intra-polyglutamine contacts and reduces the unfavorable surface energy with the surrounding solvent. As shown in Chapter 3, this unfavorable surface energy presents a large energetic barrier that prevents monomeric polyglutamine from sampling extended conformations that would be necessary for the formation of ordered β-strands. The modulation of these large energy scales is a probable mechanism by which flanking sequences alter intermolecular interactions. To see if this is the case, we fit the temperature dependence of $N^{-1}(\langle U_{total}\rangle_T - \langle U_{X17}\rangle_T) = C_1(T) + C_2(T)N^{-1/3}$ where the angular brackets denote ensemble averages. Here, N is the length of the polyglutamine segment and for a given conformation of Ac-X17-Q_N-Nme U_{total} denotes the total potential energy of the system including the milieu of counter- and co-ions; $\langle U_{X17}\rangle$ is the N independent ensemble-averaged potential energy for Ac-X17-Nme at temperature T. We assessed the temperature dependence of energetic contributions to the volumetric intra-polyglutamine interactions and the effective surface tension per residue, in the presence and absence of the X17 segment where X17 = (htt, atxn-3, atxn-6). The two energetic contributions are measured in terms of C_1(T) and C_2(T), respectively. In the poor solvent regime (T < 360 K), the presence of the X17 segment causes a loss of favorable intra-polyglutamine domain self-interactions worth ~3 kcal/mol with respect to the homopolymeric constructs, see results for C_1(T) for Q_N in Figure 5.8 panel A. This is because the globular polyglutamine domains in Ac-X17-Q_N-Nme peptides have reduced self-interactions compared to homopolymeric Ac-Q_N-Nme constructs. Panel B of Figure 5.8, which plots the

temperature dependence of $C_2(T)$ shows a decrease in effective surface tension for the globular polyglutamine domain in Ac-X17-Q_N-Nme peptides for T < 360 K. This decrease arises from the sequestration of glutamine residues in intramolecular inter-domain interfaces. As T increases, the difference between the $C_2(T)$ term for polyglutamine domains in the two constructs decreases because the inter-domain interface loosens with increasing T. The analysis in Figure 5.8 suggests that all X17 segments are gatekeeper domains because they reduce the energy penalty associated with exposing the polyglutamine domain to water by sequestering aggregation-prone polyglutamine in an inter-domain interface. Additionally, as T decreases, $C_2(T)$ decreases continually, which implies that the penalties for conformational fluctuations of the polyglutamine domain decrease substantially because of the coupling between X17 and polyglutamine. These increased conformational fluctuations are evident in Figure 5.7 by an increase in magnitude of asphericity.

The gatekeeping mechanism for these flanking sequences involves the combination of reduction in favorable intermolecular associations through sequestration of polyglutamine and electrostatic repulsions between surface-exposed charges from the X17 segment. Sequestration involves the lowering of C_1 via the introduction of new interactions that compete with polyglutamine self-interactions and a reduction in C_2 that lessens the driving force for aggregation. Atxn-3 is the most effective at disrupting polyglutamine self-interactions and reducing the surface tension over the polyglutamine segment when taken over all temperatures. This shows this N-terminal element is tightly coupled to the polyglutamine surface and when considering this segment has the highest charge density it follows that atxn-3 should be a highly effective gatekeeper. However, the energetics of atxn-6 gives us pause. Atxn-3 and atxn-6 have similar C_1 and C_2 profiles and yet atxn-3 has a significantly higher gatekeeping potential

showing that burial of polyglutamine residues alone is not a sufficient condition to inhibit intermolecular interactions. Table 5.2 provides the reason for this: polar residues dominate the sequence of atxn-6 and there is a lack of charged residues. Addition of chemically similar residues to the polyglutamine segment is equivalent to increasing the effective polyglutamine length, which in turn increases the driving force for aggregation. This has the net effect of reducing the gatekeeping potential of atxn-6.

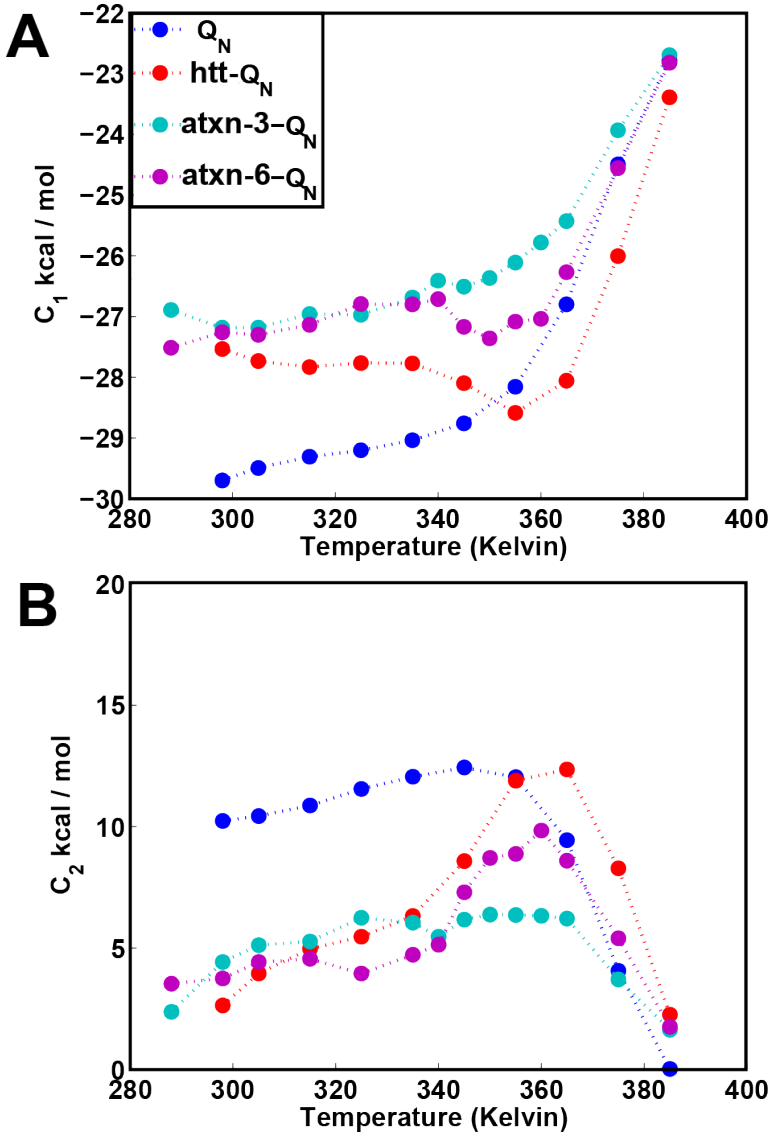


Figure 5.8: Energetics of polyglutamine self-interactions (C_1) and surface tension (C_2).

The average internal energy of the polyglutamine segment was decomposed into a volumetric intra-polyglutamine interaction term (C_1) and surface tension term (C_2). All N-terminal flanking sequences interact with the polyglutamine segment and diminish favorable polyglutamine self-interactions. They also lower the effective surface tension by sequestering exposed polyglutamine which lessens the driving force for aggregation. Panels A and B show the variation of $C_1(T)$ and $C_2(T)$, for polyglutamine domains in monomers of Ac-htt-Q_N-Nme (red), Ac-atxn-3-Q_N-Nme (cyan), Ac-atxn-6-Q_N-Nme (purple) and homopolymeric (red) constructs, respectively. The energetic terms were obtained from fits to the linear regression equation shown in the text. Fits were obtained using results for chain lengths N=(15, 20, 25, 30, 35, 45).

5.8 Intermolecular associations between polyglutamine molecules are stabilized by conformational fluctuations and molecular entanglement

Previous work [55] by Pappu lab members Andreas Vitalis and Xioling Wang quantified the thermodynamics of intermolecular associations for different temperatures and polyglutamine lengths by studying monomer-dimer equilibria. The effective peptide concentrations in these simulations match *in vitro* experiments. In the poor solvent regime the probability of forming homodimers is higher for longer chains. Chapter 3 expands upon this work and shows β-sheet formation is not a pre-requisite for stable intermolecular associations. Instead, generic parameters such as chain length and the poorness of solvent are sufficient for describing the thermodynamics of intermolecular associations. Previous work [55] also analyzed the role of spontaneous conformational fluctuations in promoting intermolecular associations. Random globular conformations were chosen from the conformational ensemble of monomeric Ac-Q_{30}-

Nme. The internal coordinates were frozen and only rigid body Monte Carlo moves were allowed for subsequent sampling. Statistics were recorded to construct the requisite histograms for intermolecular separations, R, sampled in simulations with rigid globules. The resultant average cumulative distribution was compared to that obtained for the association of fully flexible chains. These comparisons show that suppression of fluctuations causes significant diminution of intermolecular associations for Q_{30} (Figure 5.9) and this result holds for other chain lengths as well. Thus, the degree of poorness of the solvent and spontaneous fluctuations of globules make up the thermodynamic driving forces for polyglutamine aggregation.

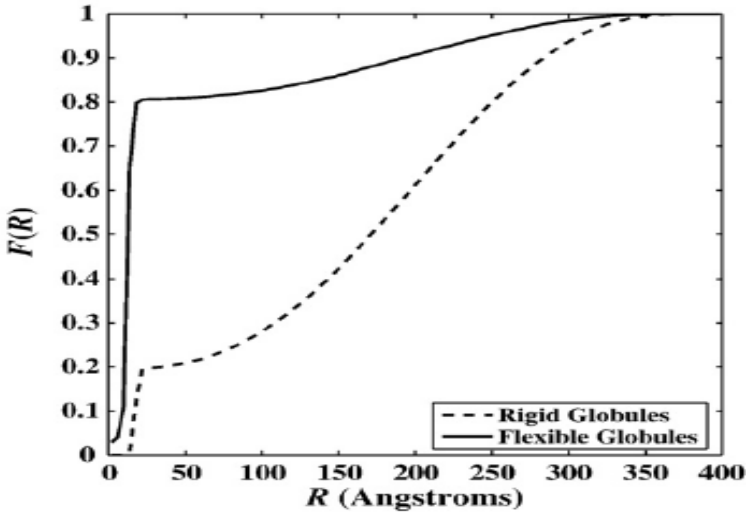


Figure 5.9: Cumulative distribution functions F(R) quantifying the probability of realizing intermolecular separation R for pairs of Q30 molecules. Results are shown for globules that are either fully flexible (solid curves) or rigid (dashed curves). This figure is reproduced from Vitalis, Wang, J. Mol. Biol (2008).

Figures 4 and 7 show evidence for significant conformational fluctuations via the departure from spherical molecular symmetry in the presence of intermolecular interactions. The magnitude of these fluctuations is dependent on flanking sequence, which implies the stability of multimolecular assemblies is flanking sequence dependent. So far, we have shown gatekeepers

inhibit aggregation via long-range electrostatic repulsions and sequestration of polyglutamine interfaces. Thus, modulation of conformational fluctuations necessary for homotypic interactions is another mechanism by which gatekeepers can elicit an effect. Figure 5.10 shows the nature of these conformational fluctuations and Figure 5.11 shows their sequence context dependence.

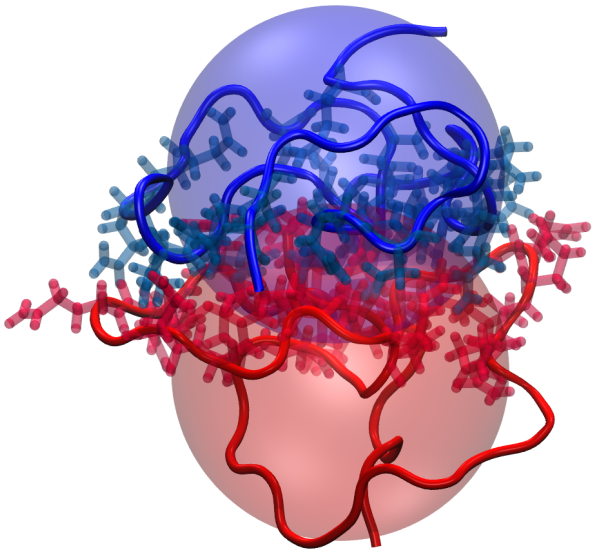


Figure 5.10: Simulation snapshot of two aggregated Ac-Q_{35}-Nme molecules with significant entanglement. The two polyglutamine chains are denoted by shades of blue and red. An opaque sphere of radius $\langle Rg \rangle_m$ is placed at the center of mass of each molecule and denotes the average radius of gyration of the monomer in isolation. The polypeptide backbone is shown as a tube and only those sidechains that participate in intermolecular contacts are shown. There is significant interpenetration of the spheres of radius $\langle Rg \rangle_m$ and sidechains from one molecule into the other. These features are present for all molecules with intermolecular interactions but the degree of interpenetration or "entanglement" varies.

Collapsed globules formed by polymers that are sufficiently long ($N \geq 15$) have solid-like cores and liquid-like surfaces [(26)]. As a result of this bipartite character, globules aggregate through contacts between their liquid-like surfaces. Radial density profiles (RDPs) are used in Lifshitz-like theories as order parameters for describing coil-to-globule transitions. These RDPs

show that the boundary between solid-like core and liquid-like surface is situated at a distance of $\langle R_g \rangle$ from the average center-of-mass of the globule [56]. As shown by Figure 5.9, reduction in the mixing between liquid-like surfaces via elimination of those degrees of freedom leads to the diminution in intermolecular interactions. This surface mixing means sidechains are entangled between molecules and this entanglement is shown in Figure 5.10. Analysis of pair correlation functions in Chapter 3 shows that entanglement involves favorable sidechain-sidechain, sidechain-backbone, and backbone-backbone contacts. The high heterogeneity of these contacts is a result of the homopolymeric nature of these molecules. Deformations of the solid-like core occur as well and are shown by the departure of each molecule from spherical symmetry and the distance between the molecular centers of mass (R) adopting values less than $2\langle Rg \rangle_m$. We define the degree of entanglement for dimers based on how far, on average, R is less than $2\langle Rg \rangle_m$. This analysis is shown in Figure 5.11 for dimers of polyglutamine with N=35 with and without flanking sequences.

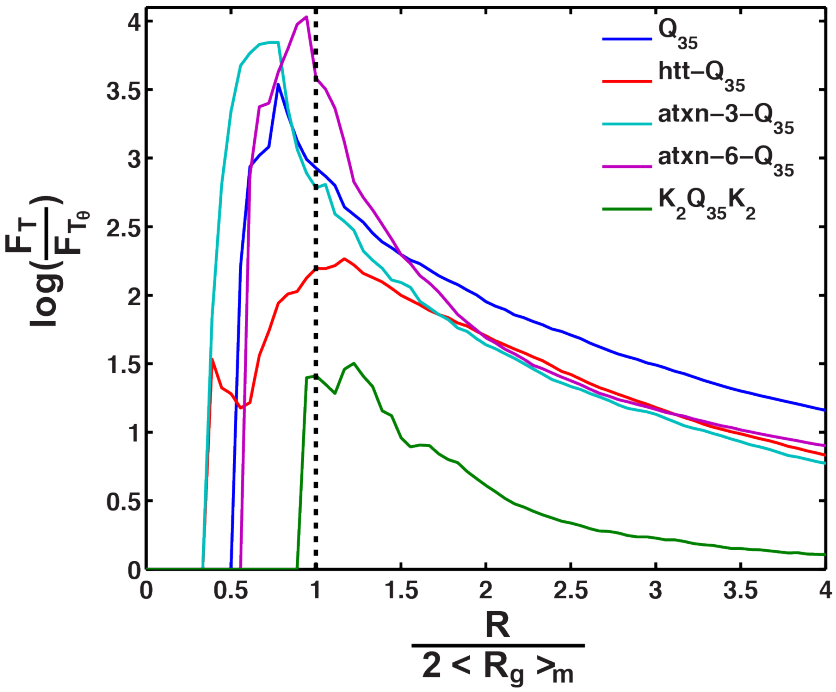


Figure 5.11: Quantification of the degree of entanglement between pairs of molecules with and without flanking sequences. Simulation results of dimers provide the normalized cumulative distributions (CDFs) of distances between centers-of-mass (R). F_T denotes the CDF collected at temperature T and $\langle R_g \rangle_m$ is the average radius of gyration of the monomer for each construct. The abscissa corresponds to R normalized by the average radii of gyration of the monomers and the ordinate plots the logarithm of the CDF at temperature T divided by this function for $T=T_\theta$. In this case, T=315K. In the absence of entanglement, these CDFs should be zero for $R \leq 2\langle R_g \rangle_m$ or there should be no density ≤ 1 for the normalized version of R denoted by the vertical dashed line. Otherwise, conformations of individual chains in aggregates (dimers) are deformed when compared to conformations of isolated globules. Clearly, the normalized CDFs have non-zero values for distances of separation that are less than 1 in terms of the normalized R, which implies significant inter-chain entanglement.

Figure 5.11 shows flanking sequences modulate two distinct length and energy scales. The first we refer to as the *docking* regime where $R / (2\langle R_g \rangle_m) \geq 1$. Here, interactions are surface mediated and rigid-body in nature. Flanking sequences that increase net charge density per monomer or increase the total number of charged residues inhibit the stability of the docking regime. The net charge per monomer / total number of charged residues for Ac-K_2-Q_{35}-K_2-Nme, Ac-atxn-3-Q_{35}-Nme, Ac-htt-Q_{35}-Nme, Ac-atxn-6-Q_{35}-Nme, Ac-Q_{35}-Nme are as follows: +4/4, +4/8, +1/5, 0/2, and 0/0 respectively. The quadruply lysinated construct is most effective at destabilizing the docking regime because long-range repulsions are found on both ends of the molecule, thus it is more difficult to find a docking orientation that is stable and does not bury charged residues. All other flanking sequences contain both positive and negative charges that can screen each other, hence long-range repulsions at length scales of $R / (2\langle R_g \rangle_m) >> 1$ are of similar magnitudes. For smaller scales relevant to the docking regime, $R / (2\langle R_g \rangle_m) \approx 1$, the flanking sequence identity and patterning of charged residues becomes important. Atxn-6 has the highest docking stability because most residues are chemically similar to glutamine. The charged termini of atxn-3 largely undocks from the polyglutamine surface. This is evident in Figure 5.8 by increased values of C_2. This allows a significant portion of the polyglutamine surface to be exposed so the docking stability is on par with that of homopolymeric polyglutamine. The htt flanking sequence remains tightly coupled to the polyglutamine surface. The presence of this charged surface coating the polyglutamine significantly decreases docking stability.

The *entangled* regime corresponds to $R / (2\langle R_g \rangle_m) < 1$. Ac-$K_2$-$Q_{35}$-$K_2$-Nme is nearly devoid of entanglement because of high and uniform charge repulsions. With the exception of htt, the stability of the entangled state is equivalent across constructs because entanglements are driven by associations between polar tracts which are present in all cases. Entanglement stability

is lower for htt because the sequestration of polyglutamine residues by htt prevents them from participating in entanglements.

5.9 Discussion and conclusions

The purpose of this chapter is to develop a quantitative understanding of how flanking sequences from disease-related proteins modulate the intrinsic, length-dependent conformational preferences and aggregation mechanisms of polyglutamine. Chapter 3 shows aggregation of homopolymeric polyglutamine is a downhill, barrier less process that does not require nucleating conformational transitions. However, this conclusion need not hold upon the introduction of flanking sequences. The results from this chapter allow us to develop a complete mechanistic framework for describing polyglutamine aggregation taking into account N-terminal flanking sequences. This framework, detailed below, is of sufficient generality that other flanking termini should be describable as well. Kiersten Ruff is currently testing this assertion and Siddique Khan, a postdoc in the Pappu lab, is testing this framework and its consequences using a coarse-grain computational model.

A generic phase diagram for a polymer solution with an upper critical solution temperature (UCST) is shown in Figure 5.12 where φ denotes the volume fraction of polymer. The binodal and spinodal refer to two distinct envelopes of points for which $(\partial \Delta G_{mix} / \partial \varphi) = 0$ and $(\partial^2 \Delta G_{mix} / \partial \varphi^2) = 0$, respectively. Here, $\Delta G_{mix}(\varphi, T)$ is the free energy of mixing. The two-phase regime where the phase separated state is thermodynamically favored lies below the binodal. The location of the binodal is defined by the saturation concentration i.e., the concentration of soluble polymer at temperature T that is at equilibrium with insoluble aggregates. The location of any point below the binodal is also a measure of its quench depth into the two-phase regime.

Nucleation: For points between the binodal and spinodal, $(\partial^2 \Delta G_{mix} / \partial \varphi^2) > 0$, the soluble phase is metastable and energy barriers must be overcome to nucleate thermodynamically favored aggregates from the homogenously mixed phase. For a fixed temperature, nucleation requires the creation of a droplet with an internal concentration larger than $|\Delta\phi|$, which is the magnitude of the gap between the binodal and spinodal. For systems with a UCST this gap increases with decreasing temperature (Figure 5.12). Bona fide nuclei should have the same composition and concentration as the phase separated state. In homogeneous nucleation, the rate of nucleation is determined by G_D, which is the barrier to diffusion of molecules between old and new phases. G_S the free energy penalty associated with forming an interface between the two phases leading to lag times for nucleation. For systems with a UCST both G_D and G_S will increase with decreasing T.

Spinodal decomposition: Below the spinodal, $(\partial^2 \Delta G_{mix} / \partial \varphi^2) < 0$, the homogeneously mixed state is unstable and there are no energy barriers for creating interfaces between phases i.e., aggregation is thermodynamically downhill. In this region, local concentration gradients engendered by conformational fluctuations provide part of the driving force for phase separation. Theory [57-59] and experiment [60, 61] have shown that the binodal and spinodal are essentially coincident in dilute solutions of generic, synthetic homopolymers.

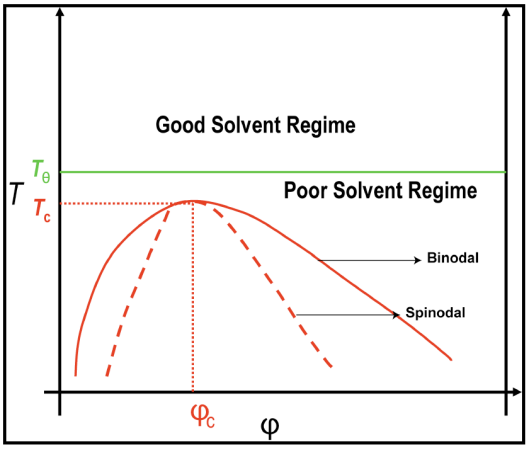


Figure 5.12: Schematic phase diagram for polymer solutions with a UCST. The point (φ_c,T_c), where the binodal and spinodal coincide, is the critical point.

For spinodal decomposition, only diffusion of individual chains or clusters of chains should limit the rate of growth of aggregates. The existence of lag times [16] would seem to rule out spinodal decomposition as a plausible mechanism for polyglutamine aggregation. However, well-reasoned analysis shows that lag times are also consistent with a variant of spinodal decomposition. In dilute solutions of properly disaggregated material, aggregation and phase separation should begin from a solution of well-dispersed globules. The time scale for collisions between globules will be $\tau_c \approx L^2/D$ where L denotes the range of interactions between globules and D is the diffusion coefficient of a globule. Aggregation requires entanglement between chains (Figure 5.9) and this degree of entanglement can be modified by flanking sequences (Figure 5.11). The time required to achieve a requisite degree of entanglement is denoted as τ_e. For $\tau_e > \tau_c$, many of the intermolecular collisions will be ineffective and the larger the gap between τ_c and τ_e the longer it takes for the necessary entanglement to be achieved. This apparent lag time has nothing to do with the presence of specific energy barriers as in classical nucleation. Instead, the mechanism of aggregation is classified as viscoelastic spinodal decomposition.

In spinodal decomposition (unlike nucleation) the rate of growth ofaggregates will increase and lag times will decrease with increased quench depths [61]. For a given concentration, longer chains will correspond to deeper quenches into the two-phase regime and published data [16, 47, 49] clearly show decreased lag times with increasing chain length. Additionally, conformational fluctuations provide the main driving force for spinodal decomposition because they create local concentration gradients. The results shown in Figures 4, 7, and 9 provide direct evidence for conformational fluctuations and poorness of solvent as combined driving forces for aggregation. Finally, indirect evidence for the separation between τ_c and τ_e comes from extrapolations of results from molecular dynamics simulations with explicit representations of water molecules [62]. Collapse is rapid and happens on the nanosecond timescale whereas conformational relaxation is considerably slower and is described by a stretched exponential of the form $\exp[-(t/\tau)^{0.15}]$, where τ is the time scale for collapse. Surface rearrangements and conversions between distinct globule conformations require timescales that are orders of magnitude longer than collapse or diffusion times of globules.

These two timescales τ_c and τ_e are modulated based on the stabilities of the *docking* and *entangled* states, respectively. The range of interactions between globules directly controls τ_c and is altered by the charge content of flanking sequences. Flanking sequences also sequester polyglutamine surfaces and this destabilizes the docking state. This means many collisions between molecules occurring on the timescale τ_c are unproductive and this widens the gap between τ_c and τ_e resulting in increased lag times. The magnitude of conformational fluctuations sets τ_e, increased fluctuations allow for faster conformational rearrangements and this decreases τ_e. As seen above, these fluctuations are tunable by flanking sequences.

We claim aggregation of polyglutamine-rich molecules follow viscoelastic spinodal decomposition, which is a family of related aggregation mechanisms. The precise details regarding the stabilities and types of intermediates as well as the kinetics of interconversion between them will be set by τ_c and τ_e. Figure 5.5 provides direct evidence for this: the rates of cluster formation and the preferred cluster sizes for synthetic polyglutamine constructs are changed by the presence of flanking charged residues and magnitude of total charge. We claim that gatekeepers are flanking sequences prevent intermolecular associations. The mechanisms of gatekeeping are: 1) long-range repulsions that minimize the overall frequencies of intermolecular encounters and confer orientational specificity to these encounters which in turn modify τ_c 2) sequestration of polyglutamine at the monomer level that in turn minimizes entanglements in favor of discretized inter-monomer interactions primarily through surface docking that alter τ_e 3) the presence of a flanking sequence that proscribes several modes of association due to large excluded volume effects. The net gatekeeping efficiency depends on a complex balance between these mechanisms and timescales τ_c and τ_e. The balance is dependent on the sequence properties of the flanking sequence. For example, both htt and atxn-6 sequester polyglutamine surfaces and yet atxn-6 has almost no gatekeeping ability. This observation follows because htt coats the surface of polyglutamine with charged residues, while atxn-6 replaces polyglutamine with other similar polar residues. This changes the stability of the docked and entangled states resulting in the difference in gatekeeping ability.

We must be clear regarding the scope of these simulation results. Monomer-dimer equilibrium is modeled by quenching the possibility of forming higher order species and this provides information regarding the implications of monomer conformations for incipient intermolecular interactions. In detail, it provides information regarding a) the conformational

changes that are needed for associations or are induced by associations b) the effects of specific flanking sequences in modulating intermolecular associations and 3) the balance between contribution of polyglutamine domains vs. flanking sequences on intermolecular association, their type, range, and strengths. However, information from such simulations does not by itself provide insights regarding the overall thermodynamic phase diagram or interpretations from kinetics experiments that quantify the formation of species that are higher order than dimer. This requires the use of either multiple molecules and/or coarse graining that builds inter-particle pair potentials using information gleaned from monomer-dimer simulations as input – as has been done by Siddique Khan, a postdoc in the Pappu lab.

The observed gatekeeping efficiencies for the flanking sequence constructs correlate with the ages of onset for their respective polyglutamine expansion diseases shown in Figure 5.1. This work is by no means exhaustive; however, continued work in the Pappu lab is underway to elucidate the effects of C-terminal sequences, the coupling between N- and C-terminal sequences, examine additional flanking sequences, and to focus on multi-molecular assemblies. The primary upshot of this work is we can use sequence properties of natural flanking sequences involved in expansion diseases to predict the effect they have on the aggregation mechanism. This insight that there are two primary energy scales, docking and entanglement, allows the mapping of flanking sequence properties to the stability of these energy scales.

5.10 Materials and methods

5.10.1 Anisotropy experiments

We used the following constructs for these measurements: K_2-Q_{40}-K_2, C*-K_2-Q_{40}-K_2, Q_{40}-K_2 and C*-Q_{40}-K_2. Here, C* denotes a cysteine residue modified with a covalently attached tetramethyl rhodamine (TMR). All peptides were disaggregated using established protocols [16].

Peptide concentrations were determined using a micro-BCA assay (Pierce, Rockford, IL) using $WQ_{40}K_2$ as the concentration standard where the concentration was determined by absorbance at 280 nm. For experiments in PBS, the unlabeled peptides were diluted into PBS buffer to a concentration of 50 μM and disaggregated TMR-labeled peptides were added to a concentration of 200 nM at a total volume of 200 μL in single wells of a Corning 96-well plate. Anisotropy was recorded in 5-minute intervals on a Tecan M100 multi-plate reader. The excitation wavelength was 533 nm with a 5 nm bandwidth and emission was monitored at 590 nm with a 10 nm bandwidth. Temperature was controlled at 25° C.

5.10.2 System setup and conformational sampling details for simulations

For monomer simulations, the following peptide constructs were used: Ac-Q_N-Nme, Ac-C-Q_N-W-Nme, Ac-K_2-C-Q_N-W-K_2-Nme, Ac-G-Q_N-C-K_2-Nme, Ac-D_2-Q_N-K_2-Nme, where Ac denotes the acetyl capping group and Nme is the N-methylamide capping group. To maintain consistency with published experimental results and asses polyglutamine length dependent properties, the following lengths of polyglutamine were used: N = (4, 7, 10, 13, 16, 24). For dimer simulations, we limited our focus to N=35 with flanking lysines.

The simulation results were generated using the ABSINTH implicit solvation model and molecular mechanics potential functions. For each peptide investigated, Table 5.3 lists the sampling methodology and number of independent sets of replicates that were used to obtain the data presented in this chapter. For monomers, a single replicate is defined as an independent simulation at a single temperature. In the case of dimers, each replicate is a single thermal replica exchange run encompassing ten different temperatures. For each monomer replica we carried out $2x10^7$ production steps after $1x10^6$ steps of equilibration while $5.15x10^7$ production and $1x10^7$ steps of equilibration were used for dimers. These settings were needed to obtain quantitatively

reliable results for the longest peptides studied. Partial charges and parameters for torsions that maintain planarity of peptide units were taken from the OPLS-AA/L forcefield. The quality of monomeric sampling was enhanced using a high number of theta state aided quenches. This involves initializing each simulation trajectory at T=298K with a random conformation drawn from another simulation performed at the theta temperature for polyglutamine T_{θ}=390K. Sampling of dimers was enhanced with thermal replica exchange using the following temperature schedule: T = 298K, 305K, 315K, 325K, 335K, 345K, 355K, 365K, 375K, 385K. Markov chain Metropolis Monte Carlo (MC) simulations were performed in the canonical ensemble using the CAMPARI software package (http://sourceforge.net/projects/campari).

Details of the MC move sets are provided in Table 5.4. The peptides were enclosed in a spherical droplet whose boundary was modeled as a stiff, one-sided harmonic potential. Note that there were a small number of moves for each dimer simulation that were used as swap attempts for replica exchange, i.e. exchange attempts were made every 50,000 MC moves. Details regarding the implementation of the individual moves along with the definitions of each column of Table 5.4 are described in Chapter 3.

For simulations of dimerization, two peptides were enclosed in a droplet of radius 200 Å to mimic an effective concentration of 100 μM. A radius of 100 Å was used for all monomer simulations. In all simulations, polymer analyses and snapshots were recorded every 5000 steps. The use of a large multiple replicates allowed us to quantify the degree of fluctuations present in the monomeric constructs.

Constructs	Sampling methodology	Quantity of MC sampling
Ac-Q_N-Nme Ac-C-Q_N-W-Nme Ac-K_2-C-Q_N-W-K_2-Nme Ac-G-Q_N-C-K_2-Nme Ac-D_2-Q_N-K_2-Nme with N=(4, 7, 10, 13, 16, 24)	T=298K 120 independent replicates	$1x10^6$ equilibration $2x10^7$ production
Ac-Q_{35}-Nme Ac-Q_{35}-K_2-Nme Ac-Q_{35}-K_2-Nme Ac-htt-Q_{35}-Nme Ac-atxn-3-Q_{35}-Nme Ac-atxn-6-Q_{35}-Nme	Thermal replica exchange at T=(298, 305, 315, 325, 335, 345, 355, 365, 375, 385) K 5 independent replicates	$1x10^7$ equilibration $5.15x10^7$ production

Table 5.3: Sampling methodology and quantity applied to each peptide system

Move type	Parameters for monomers	Parameters for dimers
Rigid-body	9% (50%, 10A, 20 deg)	27% (50%, 10A, 20 deg)
Random cluster	1% (50%, 10A, 20 deg)	3% (50%, 10A, 20 deg)
Concerted rotation	6.3%	5.6%
Omega (w)	5.67% (90%, 5 deg)	5.04% (90%, 5 deg)
Sidechain (x1, x2, x3)	27% (4x, 60%, 30 deg)	14% (4x, 60%, 30 deg)
Backbone (phi, psi)	51.03% (70%, 10 deg)	45.36% (70%, 10 deg)

Table 5.4: Overview of the frequency of the different Monte Carlo move sets used in all simulations

5.11 References

1. Williams, A. J., and Paulson, H. L. (2008) Polyglutamine neurodegeneration: protein misfolding revisited., *Trends Neurosci 31*, 521-528.
2. Ignatova, Z., and Gierasch, L. M. (2006) Extended Polyglutamine Tracts Cause Aggregation and Structural Perturbation of an Adjacent β-Barrel Protein, *Journal of Biological Chemistry 281*, 12959-12967.
3. Waelter, S., Boeddrich, A., Lurz, R., Scherzinger, E., Lueder, G., Lehrach, H., and Wanker, E. E. (2001) Accumulation of Mutant Huntingtin Fragments in Aggresome-like

Inclusion Bodies as a Result of Insufficient Protein Degradation, *Mol. Biol. Cell 12*, 1393-1407.

4. Holmberg, C. I., Staniszewski, K. E., Mensah, K. N., Matouschek, A., and Morimoto, R. I. (2004) Inefficient degradation of truncated polyglutamine proteins by the proteasome, *EMBO J 23*, 4307-4318.

5. Venkatraman, P., Wetzel, R., Tanaka, M., Nukina, N., and Goldberg, A. L. (2004) Eukaryotic Proteasomes Cannot Digest Polyglutamine Sequences and Release Them during Degradation of Polyglutamine-Containing Proteins, *Molecular Cell 14*, 95-104.

6. Sun, B., Fan, W., Balciunas, A., Cooper, J. K., Bitan, G., Steavenson, S., Denis, P. E., Young, Y., Adler, B., Daugherty, L., Manoukian, R., Elliott, G., Shen, W., Talvenheimo, J., Teplow, D. B., Haniu, M., Haldankar, R., Wypych, J., Ross, C. A., Citron, M., and Richards, W. G. (2002) Polyglutamine Repeat Length-Dependent Proteolysis of Huntingtin, *Neurobiology of Disease 11*, 111-122.

7. Thomas, P., Wilkinson, F., Man, N. T., Harper, P. S., Neal, J. W., Morris, G. E. & Jones, A. L. (1998) Full length huntingtin is not detected in intranuclear inclusions in Huntington's disease brain, *Biochem. Soc. Trans. 26*, S243.

8. Gafni, J., Hermel, E., Young, J. E., Wellington, C. L., Hayden, M. R., and Ellerby, L. M. (2004) Inhibition of Calpain Cleavage of Huntingtin Reduces Toxicity, *Journal of Biological Chemistry 279*, 20211-20220.

9. Ratovitski, T., Gucek, M., Jiang, H., Chighladze, E., Waldron, E., D'Ambola, J., Hou, Z., Liang, Y., Poirier, M. A., Hirschhorn, R. R., Graham, R., Hayden, M. R., Cole, R. N., and Ross, C. A. (2009) Mutant Huntingtin N-terminal Fragments of Specific Size Mediate

Aggregation and Toxicity in Neuronal Cells, *Journal of Biological Chemistry 284*, 10855-10867.

10. Ratovitski, T., Nakamura, M., D'Ambola, J., Chighladze, E., Liang, Y., Wang, W., Graham, R., Hayden, M. R., Borchelt, D. R., Hirschhorn, R. R. & Ross, C. A. (2007) N-terminal proteolysis of full-length mutant huntingtin in an inducible PC12 cell model of Huntington's disease, *Cell Cycle 6*, 2970-2981.
11. Walker, F. O. (2007) Huntington's disease, *The Lancet 369*, 218-228.
12. Zhang, H., Li, Q., Graham, R. K., Slow, E., Hayden, M. R., and Bezprozvanny, I. (2008) Full length mutant huntingtin is required for altered Ca2+ signaling and apoptosis of striatal neurons in the YAC mouse model of Huntington's disease, *Neurobiology of Disease 31*, 80-88.
13. Tanaka, Y., Igarashi, S., Nakamura, M., Gafni, J., Torcassi, C., Schilling, G., Crippen, D., Wood, J. D., Sawa, A., Jenkins, N. A., Copeland, N. G., Borchelt, D. R., Ross, C. A., and Ellerby, L. M. (2006) Progressive phenotype and nuclear accumulation of an amino-terminal cleavage fragment in a transgenic mouse model with inducible expression of full-length mutant huntingtin, *Neurobiology of Disease 21*, 381-391.
14. Nozaki, K., Onodera, O., Takano, H., and Tsuji, S. (2001) Amino acid sequences flanking polyglutamine stretches influence their potential for aggregate formation, *NeuroReport 12*, 3357-3364.
15. Young, J. E., Gouw, L., Propp, S., Sopher, B. L., Taylor, J., Lin, A., Hermel, E., Logvinova, A., Chen, S. F., Chen, S., Bredesen, D. E., Truant, R., Ptacek, L. J., La Spada, A. R., and Ellerby, L. M. (2007) Proteolytic Cleavage of Ataxin-7 by Caspase-7

Modulates Cellular Toxicity and Transcriptional Dysregulation, *Journal of Biological Chemistry 282*, 30150-30160.

16. Chen, S., Berthelier, V., Yang, W., and Wetzel, R. (2001) Polyglutamine aggregation behavior in vitro supports a recruitment mechanism of cytotoxicity, *Journal Of Molecular Biology 311*, 173-182.

17. Wang, X., Vitalis, A., Wyczalkowski, M. A., and Pappu, R. V. (2006) Characterizing the conformational ensemble of monomeric polyglutamine, *Proteins: Structure, Function, and Bioinformatics 63*, 297-311.

18. Vitalis, A., Wang, X., and Pappu, R. V. (2007) Quantitative Characterization of Intrinsic Disorder in Polyglutamine: Insights from Analysis Based on Polymer Theories, *93*, 1923-1937.

19. Masino, L., Kelly, G., Leonard, K., Trottier, Y., and Pastore, A. (2002) Solution structure of polyglutamine tracts in GST-polyglutamine fusion proteins, *FEBS Letters 513*, 267-272.

20. Ross, C. A., and Poirier, M. A. (2004) Protein aggregation and neurodegenerative disease., *Nat Med 10 Suppl*, S10-17.

21. Chen, S., Berthelier, V., Hamilton, J. B., O'Nuallai, B., and Wetzel, R. (2002) Amyloid-like Features of Polyglutamine Aggregates and Their Assembly Kineticsâ€ *Biochemistry 41*, 7391-7399.

22. Sikorski, P., and Atkins, E. (2004) New Model for Crystalline Polyglutamine Assemblies and Their Connection with Amyloid Fibrils, *Biomacromolecules 6*, 425-432.

23. Chen, S., Ferrone, F. A., and Wetzel, R. (2002) Huntington's disease age-of-onset linked to polyglutamine aggregation nucleation, *Proceedings of the National Academy of Sciences of the United States of America 99*, 11884-11889.

24. Scherzinger, E., Sittler, A., Schweiger, K., Heiser, V., Lurz, R., Hasenbank, R., Bates, G. P., Lehrach, H., and Wanker, E. E. (1999) Self-assembly of polyglutamine-containing huntingtin fragments into amyloid-like fibrils: Implications for Huntingtonâ€™s disease pathology, *Proceedings of the National Academy of Sciences of the United States of America 96*, 4604-4609.

25. Grosberg, A. Y., and Kuznetsov, D. V. (1992) Quantitative theory of the globule-to-coil transition. 1. Link density distribution in a globule and its radius of gyration, *Macromolecules 25*, 1970-1979.

26. Grosberg, A. Y., and Kuznetsov, D. V. (1992) Quantitative theory of the globule-to-coil transition. 2. Density-density correlation in a globule and the hydrodynamic radius of a macromolecule, *Macromolecules 25*, 1980-1990.

27. Grosberg, A. Y., and Kuznetsov, D. V. (1992) Quantitative theory of the globule-to-coil transition. 4. Comparison of theoretical results with experimental data, *Macromolecules 25*, 1996-2003.

28. Muthukumar, M. (1986) Thermodynamics of polymer solutions, *The Journal of Chemical Physics 85*, 4722-4728.

29. Fields, G. B., Alonso, D. O. V., Stigter, D., and Dill, K. A. (1992) Theory for the aggregation of proteins and copolymers, *The Journal of Physical Chemistry 96*, 3974-3981.

30. Raos, G., and Allegra, G. (1996) Chain collapse and phase separation in poor-solvent polymer solutions: A unified molecular description, *The Journal of Chemical Physics 104*, 1626-1645.

31. Raos, G., and Allegra, G. (1996) A Cluster of Chains Can Be Smaller Than a Single Chain: New Interpretation of Kinetics of Collapse Experiments, *Macromolecules 29*, 8565-8567.

32. Raos, G., and Allegra, G. (1997) Macromolecular clusters in poor-solvent polymer solutions, *The Journal of Chemical Physics 107*, 6479-6490.

33. Nishio, I., Sun, S.-T., Swislow, G., and Tanaka, T. (1979) First observation of the coil-globule transition in a single polymer chain, *Nature 281*, 208-209.

34. Nishio, I., Swislow, G., Sun, S.-T., and Tanaka, T. (1982) Critical density fluctuations within a single polymer chain, *Nature 300*, 243-244.

35. Sun, S.-T., Nishio, I., Swislow, G., and Tanaka, T. (1980) The coil--globule transition: Radius of gyration of polystyrene in cyclohexane, *The Journal of Chemical Physics 73*, 5971-5975.

36. Crick, S. L., Jayaraman, M., Frieden, C., Wetzel, R., and Pappu, R. V. (2006) Fluorescence correlation spectroscopy shows that monomeric polyglutamine molecules form collapsed structures in aqueous solutions, *Proceedings of the National Academy of Sciences 103*, 16764-16769.

37. Vitalis, A., Wang, X., and Pappu, R. V. (2008) Atomistic Simulations of the Effects of Polyglutamine Chain Length and Solvent Quality on Conformational Equilibria and Spontaneous Homodimerization, *Journal of Molecular Biology 384*, 279-297.

38. Haacke, A., Broadley, S. A., Boteva, R., Tzvetkov, N., Hartl, F. U., and Breuer, P. (2006) Proteolytic cleavage of polyglutamine-expanded ataxin-3 is critical for aggregation and sequestration of non-expanded ataxin-3, *Hum. Mol. Genet. 15*, 555-568.

39. Bhattacharyya, A., Thakur, A. K., Chellgren, V. M., Thiagarajan, G., Williams, A. D., Chellgren, B. W., Creamer, T. P., and Wetzel, R. (2006) Oligoproline Effects on Polyglutamine Conformation and Aggregation, *Journal of Molecular Biology 355*, 524-535.

40. Thakur, A. K., Jayaraman, M., Mishra, R., Thakur, M., Chellgren, V. M., L Byeon, I.-J., Anjum, D. H., Kodali, R., Creamer, T. P., Conway, J. F., M Gronenborn, A., and Wetzel, R. (2009) Polyglutamine disruption of the huntingtin exon 1 N terminus triggers a complex aggregation mechanism, *Nat Struct Mol Biol 16*, 380-389.

41. Tam, S., Spiess, C., Auyeung, W., Joachimiak, L., Chen, B., Poirier, M. A., and Frydman, J. (2009) The chaperonin TRiC blocks a huntingtin sequence element that promotes the conformational switch to aggregation, *Nat Struct Mol Biol 16*, 1279-1285.

42. Walters, R. H., and Murphy, R. M. (2009) Examining Polyglutamine Peptide Length: A Connection between Collapsed Conformations and Increased Aggregation, *Journal Of Molecular Biology 393*, 978-992.

43. Takahashi, T., Kikuchi, S., Katada, S., Nagai, Y., Nishizawa, M., and Onodera, O. (2007) Soluble polyglutamine oligomers formed prior to inclusion body formation are cytotoxic, *Hum. Mol. Genet.*, ddm311.

44. Alberti, S., Halfmann, R., King, O., Kapila, A., and Lindquist, S. (2009) A Systematic Survey Identifies Prions and Illuminates Sequence Features of Prionogenic Proteins, *137*, 146-158.

45. Williamson, T. E., Vitalis, A., Crick, S. L., and Pappu, R. V. (2009) Modulation of Polyglutamine Conformations and Dimer Formation by the N-Terminus of Huntingtin, *Journal of Molecular Biology 396*, 1295-1309.

46. Lee, C. C., Walters, R. H., and Murphy, R. M. (2007) Reconsidering the mechanism of polyglutamine peptide aggregation., *Biochemistry 46*, 12810-12820.

47. Chen, S. M., Berthelier, V., Hamilton, J. B., O'Nuallain, B., and Wetzel, R. (2002) Amyloid-like features of polyglutamine aggregates and their assembly kinetics, *Biochemistry 41*, 7391-7399.

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

49. Chen, S. M., Ferrone, F. A., and Wetzel, R. (2002) Huntington's disease age-of-onset linked to polyglutamine aggregation nucleation, *Proceedings Of The National Academy Of Sciences Of The United States Of America 99*, 11884-11889.

50. Perutz, M. F., Pope, B. J., Owen, D., Wanker, E. E., and Scherzinger, E. (2002) Aggregation of proteins with expanded glutamine and alanine repeats of the glutamine-rich and asparagine-rich domains of Sup35 and of the amyloid beta-peptide of amyloid plaques, *Proc Natl Acad Sci U S A 99*, 5596-5600.

51. Crick, S. L., Jayaraman, M., Frieden, C., Wetzel, R., and Pappu, R. V. (2006) Fluorescence correlation spectroscopy shows that monomeric polyglutamine molecules form collapsed structures in aqueous solutions, *Proceedings Of The National Academy Of Sciences Of The United States Of America 103*, 16764-16769.

52. Fiumara, F., Fioriti, L., Kandel, E. R., and Hendrickson, W. A. (2010) Essential Role of Coiled Coils for Aggregation and Activity of Q/N-Rich Prions and PolyQ Proteins, *Cell 143*, 1121-1135.

53. MacCallum, B. (2004) Order/Disorder Prediction for Protein Sequences.

54. Williamson, T. E., Vitalis, A., Crick, S. L., and Pappu, R. V. (2010) Modulation of Polyglutamine Conformations and Dimer Formation by the N-Terminus of Huntingtin, *Journal of Molecular Biology 396*, 1295-1309.

55. Vitalis, A., Wang, X., and Pappu, R. V. (2008) Atomistic simulations of the effects of polyglutamine chain length and solvent quality on conformational equilibria and spontaneous homodimerization., *J Mol Biol 384*, 279-297.

56. Alexei R Khokhlov, A. Y. G., Vijay S. Pande (1994) *Statistical Physics of Macromolecules*, American Institute of Physics, Woodbury, NY.

57. Raos, G., and Allegra, G. (1996) Chain interactions in poor-solvent polymer solutions: Equilibrium and nonequilibrium aspects, *Macromolecules 29*, 6663-6670.

58. Raos, G., and Allegra, G. (1996) Chain collapse and phase separation in poor-solvent polymer solutions: A unified molecular description, *Journal Of Chemical Physics 104*, 1626-1645.

59. Raos, G., and Allegra, G. (1997) Macromolecular clusters in poor-solvent polymer solutions, *Journal Of Chemical Physics 107*, 6479-6490.

60. Zhang, G. Z., and Wu, C. (2006) Folding and formation of mesoglobules in dilute copolymer solutions, *Advances in Polymer Science 195*, 101-176.

61. Piçarra, S., and Martinho, J. M. G. (2000) Viscoelastic Effects on Dilute Polymer Solutions Phase Demixing: Fluorescence Study of a Poly(ε-caprolactone) Chain in THF, *Macromolecules 34*, 53-58.

62. Vitalis, A., Wang, X., and Pappu, R. V. (2007) Quantitative characterization of intrinsic disorder in polyglutamine: insights from analysis based on polymer theories., *Biophys J 93*, 1923-1937.

Chapter 6

A quantitative measure for protein conformational heterogeneity

6.1 Introduction

Proteins undergo disorder-to-order transitions either as units that fold autonomously[1] or as intrinsically disordered proteins (IDPs)[2] that couple their folding to binding[3] or self-assembly[4]. The driving forces for and mechanisms of disorder-to-order transitions are governed by the degree of conformational heterogeneity within disordered states and the extent of overlap between conformational ensembles of disordered and ordered states. Therefore, there is growing interest in quantitative studies of disordered states of proteins [5-8].

Studies of disorder in protein folding are focused on characterizing the ensemble of non-native conformations under denaturing as well as native conditions [9-11]. Of interest are questions pertaining to the degree of conformational heterogeneity[12, 13], the balance between intrachain and chain-solvent interactions that define polymeric properties[14-16], effects of macromolecular crowding[17, 18], intermolecular interactions that lead to protein aggregation[19-21], and the timescales for conversion between distinct conformations that contribute to internal friction[22]. Recent interest has also focused on the topic of IDPs. Their sequences encode preferences for heterogeneous ensembles of conformations as the thermodynamic ground state under standard physiological conditions (aqueous solutions, 150 mM monovalent salt, low concentrations of divalent ions, pH 7.0, and temperature in the 25ºC – 37ºC range)[23, 24]. Conformational heterogeneity of IDPs in their unbound forms influences their ability to adopt different folds in the context of binary and multimolecular complexes[25, 26]. In IDPs, disorder-to-order transitions

are realized by coupling the folding process to either binding or self-assembly providing the heterotypic or homotypic interactions in *trans* can stabilize the IDP in a specific fold. The stabilities of complexes are thermodynamically linked to the ensemble of conformations that IDPs sample as autonomous units.

Thermodynamic descriptions of disorder to order transitions require the use of a suitable order parameter. A *bona fide* order parameter has to quantify the symmetry that is broken as a result of the disorder-to-order transition. Proteins are polymers and can expand to form expanded, low-density conformations that have large interfaces with the surrounding solvent or collapsed high-density conformations that minimize the chain-solvent interface. It is well established that $s^2 = \frac{\langle R_g^2 \rangle}{N}$ is a *bona fide* order parameter for quantifying density changes that accompany coil-to-globule transitions[27, 28]. Here, R_g denotes the radius of gyration and N is the chain length. In protein folding, changes in density are also associated with the acquisition of an ensemble of self-similar conformations and therefore s^2 can be used as the sole parameter to monitor folding if and only if proteins follow two-state behavior[29]. Theories, simulations, and experiments have established that while s^2 is extremely important for understanding the convolution of coil-to-globule transitions with protein folding, it is inadequate for providing a complete description of transitions between unfolded and folded states[30-37]. Recent simulations and experiments have also shown that several IDPs undergo collapse to form globules under standard physiological conditions[38-44]. This preference for globules can be reversed through increases in temperature[45], net charge per residue[46, 47], or concentrations of chemical denaturants[39]. Collapse in globule forming IDPs does not have to imply the acquisition of an ensemble of self-similar

conformations. These results highlight the need for additional parameters that report on overall conformational heterogeneity.

Figure 6.1 summarizes the temperature dependence of densities and density fluctuations that are obtained from atomistic simulation results for five archetypal systems. Details of the simulations that were used to generate the temperature dependent profiles are discussed in the methods section. The N-terminal domain of the ribosomal L9 protein (NTL9) and the B1 domain of protein G (GB1) undergo unfolding transitions as temperature increases. This unfolding is also associated with chain expansion as shown in panels (A) and (B) of Figure 6.1. We compare these profiles to the temperature dependence of s^2 and density (ρ) for three homopolymeric systems. These are polyproline (P_{56}), which is intrinsically stiff, polyarginine (R_{56}), which is a highly charged polyelectrolyte, and polyglutamine (Q_{56}), which is an intrinsically disordered polar tract. P_{56} shows weak chain contraction as temperature increases. This feature is consistent with the so-called inverse transition temperature[48] that has been observed for poly-L-proline polymers and derives, partially, from the increased fraction of *cis* peptide bonds at higher temperatures[49]. R_{56} and Q_{56} show distinct limiting behaviors; the former maintains its swollen random coil behavior across all temperatures whereas the latter undergoes a globule-to-coil transition as temperature increases. Despite undergoing reversible coil-to-globule transitions, previous simulations and experimental studies demonstrate that collapse does not imply the acquisition of an ensemble of self-similar conformations *i.e.*, collapse does not imply folding[50, 51]. This in turn implies that while the temperature dependence of s^2 provides information regarding coil-to-globule transitions, it fails to distinguish between systems such as NTL9 and GB1 on the one hand and Q_{56} on the other. Analysis of the temperature dependence of the specific heat capacities (panel D in Figure 6.1), which reports on the temperature dependence of the energy variance, does not

provide any additional information that cannot be obtained by analyzing the temperature dependence of density fluctuations.

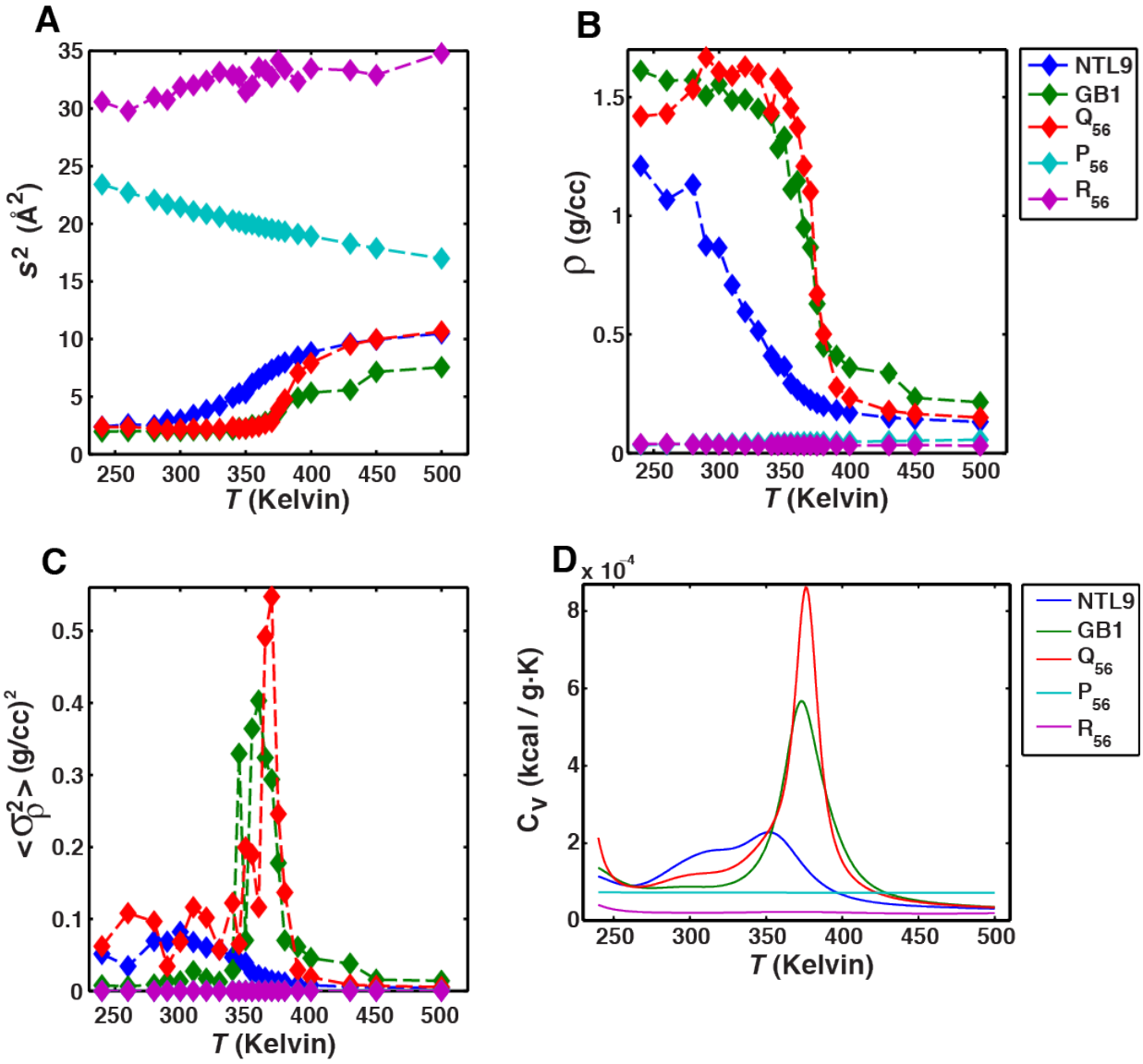


Figure 6.1: Temperature dependence of R_g, density, as well as fluctuations in density and energy for archetypes of intrinsically disordered systems and folded proteins. Each panel shows results for five different systems. Panel B quantifies the temperature dependence of chain density (in units of gm/cm^3), which is calculated as $\langle \rho \rangle = \frac{\mathrm{MW}}{\langle R_g^2 \rangle^{\frac{3}{2}}}$, where MW denotes the

molecular weight in gm-mol^{-1}. Systems undergoing a globule-to-coil transition as a function of increased temperature show a decrease in density as they transition from the high-density globule to the low-density coil. The small decrease in R_g as a function of increased temperature for P_{56} translates into negligible change in density. Panel C shows the temperature dependence of the density fluctuations for each system. This is quantified as the variance of the density distribution for a given temperature *i.e.*, $\sigma_\rho^2 = \langle \rho^2 \rangle - \langle \rho \rangle^2$. Finally, panel D quantifies the temperature dependence of the specific heat capacity. Typically, one expects sharp transitions for well-defined order-to-disorder transitions and yet, interestingly, the Q_{56} system shows the sharpest transition. The relatively broad transitions for NTL9 and GB1 highlight the joint contributions of gradual melting and different degrees of residual local structure in their unfolded states. The specific, constant volume heat capacities were calculated as $C_V = \frac{1}{\mathrm{MW}}\left(\frac{\partial \langle E \rangle}{\partial T}\right)_V$ where MW is the molecular weight and $\langle E \rangle$ is the ensemble-averaged potential energy for simulated ensembles at a given temperature.

In the parlance of energy landscape theory[52-54], a system such as polyglutamine has a rugged landscape below its collapse transition temperature[55]. Indeed, such a scenario has been predicted for IDPs[56, 57] and random polypeptide sequences[58, 59] (see Figure 6.2). This ruggedness is not registered in measures such as estimates of density or energy fluctuations because distinct conformations of equivalent compactness have negligible energy differences and hence equivalent likelihoods of being accessed. In this scenario, both the energy and density fluctuations will be small and the sharpness of the change in energy and density fluctuations masks the fact that the globule-to-coil transition in a system like polyglutamine might actually be

a disorder-to-disorder transition where the transition is between distinct classes of heterogeneous conformational ensembles.

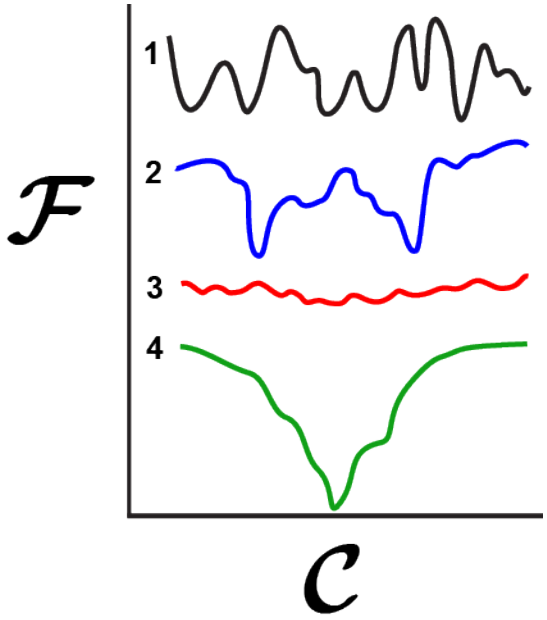


Figure 6.2: Schematic depicting different categories of free energy landscapes for different sequence- encoded degrees of conformational heterogeneity. The landscapes are shown as one-dimensional projections along a collective conformational coordinate labeled $\mathcal{C}$ and we stipulate that the temperatures are equivalent for all four cases. Schematic 1 depicts a rugged "egg-carton" landscape. This situation corresponds to globule-forming sequences that cannot distinguish between large numbers of energetically equivalent, albeit distinct conformations that all have equivalent compactness. Free energy minima are depicted as traps separated by barriers. The energy scales are such that $\Delta\mathcal{F}_M$, the average free energy difference between minima ~kT, whereas the $\Delta\mathcal{F}_B$, the average free energy difference between a barrier and adjacent local minimum is considerably larger than kT. Here, k denotes Boltzmann's constant and T is the temperature. Schematic 2 depicts discrete disorder where the system has access to a small number of distinct free energy minima that are separated by large barriers. In this case, $\Delta\mathcal{F}_M \approx$

$\Delta\mathcal{F}_B$ and both energy scales are considerably larger than kT. Schematic 3 depicts the free energy landscape for proteins that are akin to generic random-coils. Here, $\Delta\mathcal{F}_M \approx \Delta\mathcal{F}_B$, although unlike panel A, both energy scales are of order kT or smaller. Finally, schematic 4 depicts the strongly funneled landscape, which is expected for proteins that fold into well-defined three-dimensional structures.

In order to detect putative disorder-to-disorder transitions that are masked when analyzing s^2 or densities, we need a measure of heterogeneity within a conformational ensemble. For this we introduce a parameter Φ whose design is guided by the need to distinguish between systems that strongly couple their unfolding-folding transitions with coil-to-globule transitions and those systems that undergo coil-to-globule transitions with no evidence of acquiring an ensemble of self-similar conformations upon collapse. The design of Φ is also intended to accomplish two additional goals for the analysis of results from molecular simulations: (i) To compare the degree of conformational heterogeneity of ensembles obtained for a specific system at different simulation temperatures; and (ii) To compare the degree of conformational heterogeneity of ensembles for different polypeptide sequences at equivalent simulation temperatures.

The remainder of the narrative is organized as follows: In the methods section we summarize the simulation approach used to generate the conformational ensembles that were used to prototype Φ. The results section is split into two parts. In part 1, we describe the methodological framework for calculating Φ. In doing so, we discuss the choices made in converging upon the overall approach. In part 2, we use Φ to assess recent simulation results that were reported for the basic regions (bRs) of bZIP transcription factors[60]. These published results demonstrated the role of sequence contexts in modulating the intrinsic helicities of bZIP-bRs.

We show that Φ unmasks the weaknesses inherent to measures of average secondary structure contents as probes for structure. We highlight how conformational heterogeneity can prevail in ensembles with high average helicities. We conclude with a discussion section that summarizes the uses for Φ in analyzing protein disorder and in *de novo* sequence design. The discussion also provides a comparison between Φ and other approaches for quantifying conformational heterogeneity.

6.2 Methods

6.2.1 Polypeptide systems included in this work

We simulated homopolymers of glutamine (Q_{56}), proline (P_{56}), and arginine (R_{56}) each 56-residues long. In addition, we included two 56-residue polypeptides, NTL9 and GB1 that adopt well-defined folds at low temperatures. Homopolymers were N-terminally acetylated and C-terminally N′-methylamidated. Atomistic Markov Chain Metropolis Monte Carlo (MC) simulations[(61)] were performed in the canonical ensemble using one polypeptide for each construct. Mobile sodium and chloride ions were included for peptides containing charged residues and the ions were represented explicitly. The salt concentration of ion-containing systems was set to 25 mM. The peptides and ions were enclosed in a spherical droplet of radius 200 Å and the droplet boundary was enforced using a stiff harmonic boundary potential. For the basic region leucine zipper transcription factor (bZIP-bR) peptides we analyzed simulation results from the work of Das et al.

6.2.2 Details of the Metropolis Monte Carlo (MC) simulations

The CAMPARI molecular simulation package (http://campari.sourceforge.net/) in conjunction with the ABSINTH implicit solvation model[62] and OPLS-AA/L[63] molecular mechanics force field parameters (abs3.2_opls.prm) were used for all simulations. The spatial cutoffs for Lennard–Jones and electrostatic interactions between net-neutral charge groups were set to 10 Å and 14 Å, respectively. No cutoffs were employed for computing the electrostatic interactions for ions and sidechain moieties with a net charge. Sodium and chloride ions were modeled explicitly and polypeptides were modeled in atomic detail. The internal degrees of freedom included the backbone ϕ, ψ, ω and sidechain χ dihedral angles. Rigid-body moves simultaneously change rotational and translational degrees of freedom of the protein whereas translational moves were applied to alter the positions of mobile ions. Random cluster moves alter the rigid body coordinates of multiple molecules at once. Pucker moves perturb the ring geometry of proline residues[49]. The frequencies with which different moves were chosen along with parameters specific to each move type are summarized in the decision tree shown in Figure 6.3. The starting conformation for homopolymers Q_{56}, P_{56}, and R_{56} was generated at random from a pre-equilibrated distribution of atomistic self-avoiding random walks. Starting conformations for NTL9 and GB1 for all simulation temperatures were derived from Protein Data Bank (http://www.rcsb.org) IDs 2HBB and 1GB1, respectively. Additional details regarding the setup of the initial folded conformations are as described in Meng et al.[14] The bond lengths and bond angles were fixed at values prescribed by Engh and Huber[64].

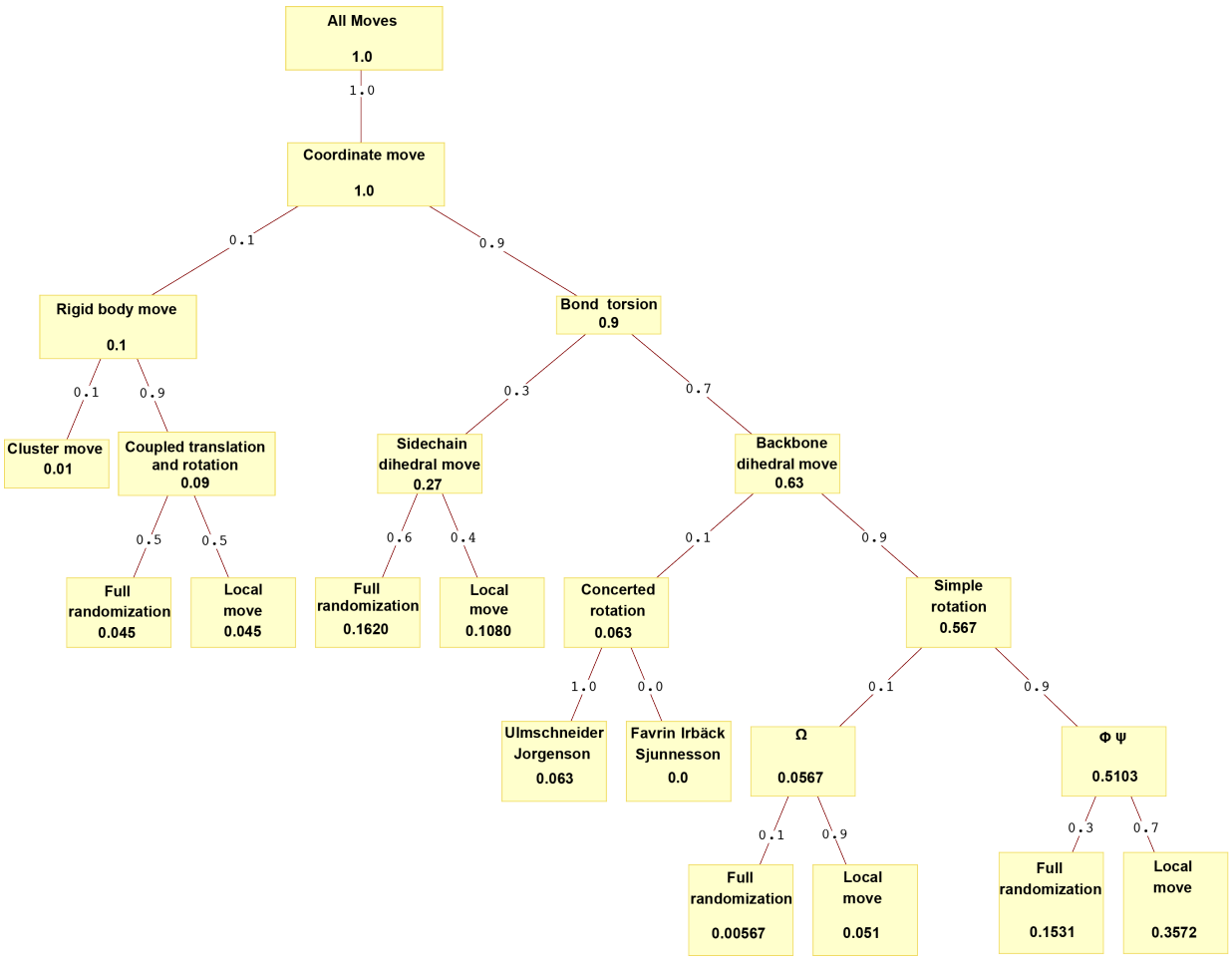


Figure 6.3: The decision tree used to select an MC move at each step. Each non-leaf node corresponds to a class of moves; each node is annotated with the overall probability of that move or class of moves being selected. Each edge is annotated with the probability of the decision process branching towards the child once the parent has been reached. The decision is complete once a leaf node is reached. Here we show the decision tree used for designing move sets in simulations of the R_{56}, NTL9, and GB1 systems.

6.2.3 The MC sampling protocol

Simulation results for NTL9, GB1, Q_{56}, P_{56}, and R_{56} were generated using the following protocol: For each system we performed ten independent MC simulations at each of the

following temperatures: T = 240, 260, 280, 290, 300, 310, 320, 330, 340, 345, 350, 355, 360, 365, 370, 375, 380, 390, 400, 430, 450, 500 K. Each independent simulation used a different random seed to initialize the MC run. A total of 8×10^7 MC steps were used in each independent simulation and of these, the results from the first 2×10^7 steps were discarded as equilibration. Observables were accumulated every 10^4 MC steps and conformational vectors for the heterogeneity calculation were collected every 10^5 steps. Thus, an ensemble from a single run that was used to calculate Φ contained 6000 members for each temperature. Reproducibility of the simulation results across multiple independent runs negated the need for using enhanced sampling methods.

6.2.4 The Flory Random Coil (FRC) model[(65)]

The FRC reference state was constructed for each polypeptide. FRC peptides were represented in all atom detail with the same degrees of freedom as used in the MC simulations described above. FRC conformations were generated by random assignment of sterically allowed combinations of backbone ϕ, ψ, ω and sidechain χ dihedral angles while ignoring all inter-residue interactions. Each step of FRC sampling consisted of picking a residue at random then assigning all of the torsional degrees of freedom of the residue to a vector of ϕ, ψ, ω and χ selected at random from a library of size 10^4. These libraries were generated for each residue by MC simulations of the corresponding dipeptides in the excluded volume (EV) limit. EV ensembles were generated using atomistic descriptions of the dipeptide while ignoring all non-bonded interactions excepting steric repulsions. A total of 4×10^7 steps were applied in each FRC simulation and resulting polypeptide conformations were accumulated every 10^5 steps. Ten independent FRC simulations were performed resulting in a total of 4000 reference conformations for each peptide. This pool of conformations is referred to as the FRC ensemble

and all members were used in calculating **Φ**.

6.3 Results

6.3.1 Estimating Φ

Our goal is to quantify the degree of conformational heterogeneity given an ensemble of conformations. This requires a method to quantify the degree of similarity between all distinct pairs of conformations within the ensemble. The resultant distribution of pairwise similarity measures is then used to obtain a value for Φ that reports on the degree of conformational heterogeneity within the ensemble. For a chain of N residues, each conformation c is represented as an $n_d \times 1$ conformational vector $\mathbf{V}_c$ where $n_d = \frac{N(N-1)}{2}$, $\mathbf{V}_c = \{d_{12}, d_{13}, \ldots, d_{N-1,N}\}$, and each element d_{ij} in $\mathbf{V}_c$ represents the spatial distance between a unique pair of residues, i and j. For each pair of residues i and j, we calculate $d_{ij} = \frac{1}{Z_{ij}} \cdot \sum_{m \in i} \sum_{n \in j} \left| \mathbf{r}_m^i - \mathbf{r}_n^j \right|$. Here, $\mathbf{r}_m^i$ and $\mathbf{r}_n^j$ denote the position vectors of atoms m and n within residues i and j, respectively, and Z_{ij} is the number of unique pairwise inter-atomic distances between the two residues.

To compare a pair of conformations k and l, we calculate a pairwise dissimilarity measure $\mathcal{D}_{kl} = 1 - \cos(\Omega_{kl})$ where $\cos(\Omega_{kl}) = \frac{\mathbf{V}_k \cdot \mathbf{V}_l}{|\mathbf{V}_k||\mathbf{V}_l|}$. An ensemble of conformations produces an ensemble of conformational vectors, $\mathbf{V}_1$, $\mathbf{V}_2$,…etc. These vectors are used to calculate a distribution of conformational dissimilarity values $P(\mathcal{D})$. Examples of these distributions are shown in Figure 6.4. For a given simulation temperature T, the first moment of the distribution of dissimilarity values denoted as $\langle \mathcal{D} \rangle_T$ provides a suitable measure of the degree of heterogeneity

within the ensemble. This measure, however, needs calibration in order for it to be used for comparing ensembles across different simulation temperatures or ensembles for different systems. For a given system, the intrinsic conformational properties of amino acids within the sequence place an upper bound on the degree of dissimilarity that is realizable[65]. For example, the intrinsic flexibility of a proline-rich sequence is inherently different from the intrinsic flexibility of a glycine-rich sequence. These differences need to be accounted for and normalized against if we are to compare the degree of conformational heterogeneity between systems. Furthermore, considerations of chain connectivity might render many of the values for inter-residue distances d_{ij} to either be invariant or slowly varying as temperature changes. These minor changes to d_{ij} can mask the true dissimilarity between pairs of conformations. Accordingly, we use a simulation approximation of the Flory random coil (FRC) that is based on the rotational isomeric approximation to calibrate the distribution of dissimilarity values obtained for ensembles of a given system. This is accomplished by calculating the pairwise conformational dissimilarity $\mathcal{D}_{kl}$ between each conformation k from the ensemble of interest and conformation l drawn from the ensemble of FRC conformations. The latter ensemble varies depending on the amino acid sequence and remains invariant with temperature. Consequently, for each ensemble corresponding to a given simulation temperature T, we obtain two distributions of dissimilarities *viz*., the distribution of $\mathcal{D}$-values for pairs of conformations within an ensemble and a distribution of $\mathcal{D}$-values comparing each conformation to an ensemble of FRC conformations (see Figure 6.4). Averaging over the former yields $\langle\mathcal{D}\rangle_T$ and averaging over the latter, which is an ensemble of ensembles yields $\langle\langle\mathcal{D}\rangle\rangle_{\text{FRC}}$. The values of $\langle\mathcal{D}\rangle_T$ and $\langle\langle\mathcal{D}\rangle\rangle_{\text{FRC}}$ lead to an estimate of the degree of heterogeneity Φ_T within the ensemble at temperature T. We first compute the ratio $\mathcal{H}_T = \langle\mathcal{D}\rangle_T / \langle\langle\mathcal{D}\rangle\rangle_{\text{FRC}}$ and use it to calculate Φ_T using $\Phi_T = 1 - \mathcal{H}_T$.

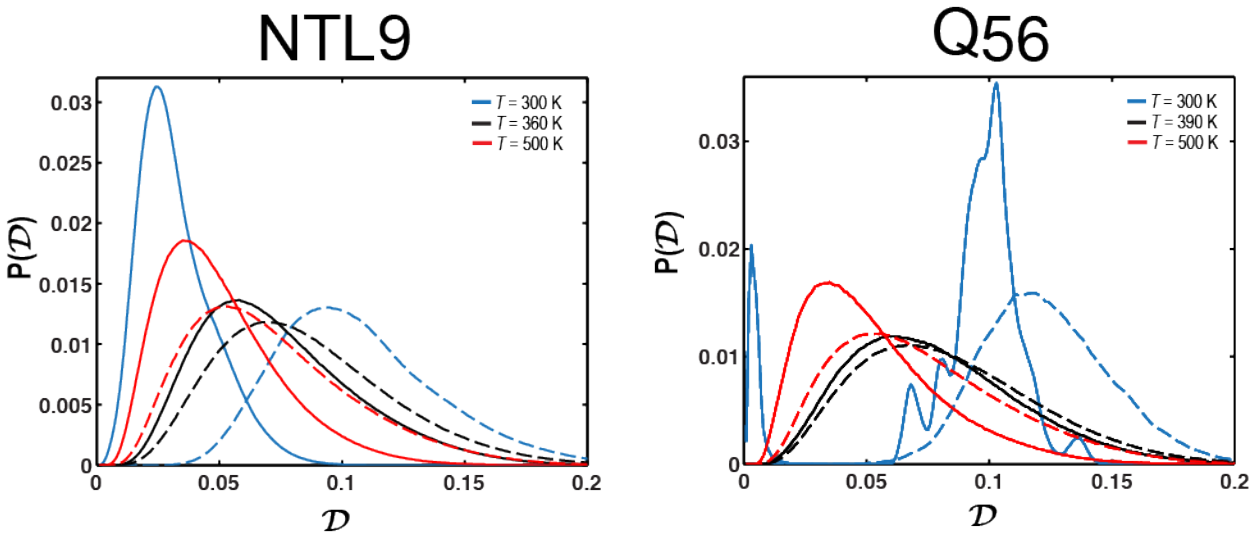


Figure 6.4: Sample distributions $P(\mathcal{D})$ for two systems at different temperatures. The panel on the left shows $P(\mathcal{D})$ distributions for NTL9 at three different temperatures. The solid curves represent intra-ensemble $P(\mathcal{D})$ distributions whereas the dashed curves are for comparisons between conformations within an ensemble at temperature T and conformations drawn from the FRC ensemble. The panel on the right shows two sets of similar distributions for the Q_{56} system at three different simulation temperatures.

6.3.2 Properties of Φ_T

The parameter Φ_T is bounded *i.e.*, $0 \leq \Phi_T \leq 1$. This property obtains because $\mathcal{H}_T \leq 1$ and results from the construction of the reference FRC ensemble, which ensures that $\langle \mathcal{D} \rangle_T \leq \langle\langle \mathcal{D} \rangle\rangle_{\text{FRC}}$. In the FRC model, the conformational partition function for the polypeptide is written as a product of partition functions of independent interaction units. All interactions between non-nearest neighbor residues are ignored while the intrinsic conformational preferences of individual residues are captured in terms of weights for each of the possible rotational isomers. The FRC ensemble therefore represents an intuitive upper bound on conformational

heterogeneity. If the degree of intra-ensemble conformational heterogeneity is akin to the upper bound on heterogeneity expected for an FRC ensemble, then the ratio $\mathcal{H}_T \to 1$ and $\Phi_T \to 0$, indicating a maximally heterogeneous ensemble. Conversely, for an ensemble of self-similar conformations it follows that, $\mathcal{D}_T \to 0$, $\mathcal{H}_T \to 0$, and $\Phi_T \to 1$.

6.3.3 Assessment of conformational ensembles using Φ_T

Figure 6.5 shows the variation of Φ_T with temperature for each of the five systems that were introduced in Figure 6.1. For NTL9 and GB1 the unfolding transition is manifest as a transition between a high value for Φ_T at low temperatures and a low value for Φ_T at high temperatures and the transition between these two limits is sharp. The slope of transition region quantifies the "rate" of change in the degree of conformational heterogeneity with temperature. In contrast to NTL9 and GB1, the temperature dependence of Φ_T for Q_{56} is consistent with equivalent degrees of heterogeneity in the high and low temperature regimes. Previous work on polyglutamine led to an estimate of $T_\theta \approx 390$ K for the theta temperature. At $T \approx T_\theta$, chain-chain and chain-solvent interactions are counterbalanced and conformational properties at the theta temperature resemble that of the FRC model. Accordingly, the temperature dependence of Φ_T for Q_{56} shows a dip and approaches zero near T_θ. Apart from this deviation, the profile of Φ_T for Q_{56} is consistent with the hypothesis of a disorder-to-disorder transition. When combined with the analysis of s^2 or ρ, it becomes clear that the polyglutamine system transitions between two classes of disorder *viz*., a heterogeneous ensemble of compact conformations that maximize the density at low temperatures and a heterogeneous ensemble of expanded conformations that minimize the density at high temperatures.

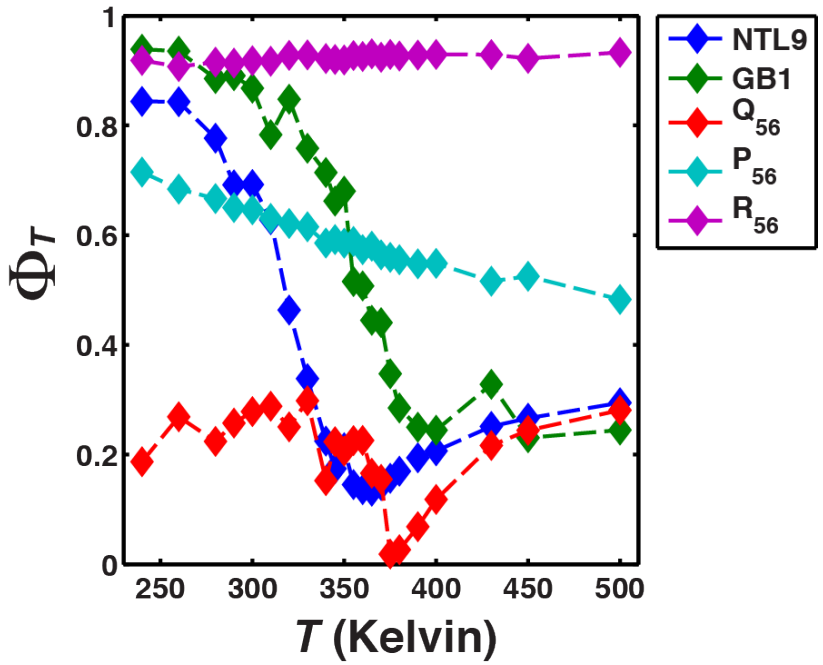


Figure 6.5: Temperature dependence of Φ for the five archetypal systems. For systems that undergo globule-to-coil transitions, the values of Φ are lowest in the vicinity of the theta temperature. This temperature varies based on sequence composition. The fractal nature remains invariant for the polyelectrolyte R_{56}, and consequently Φ remains uniformly high for this system. While Q_{56} goes through a globule-to-coil transition the degree of conformational heterogeneity, as measured by Φ, does not change with temperature, which suggests that this globule-to-coil transition is an example of a disorder-to-disorder transition. NTL9 and GB1 undergo globule-to-coil transitions and these are accompanied by order-to-disorder transitions as shown by the respective profiles for Φ.

The Φ_T profiles for NTL9 and GB1 also show dips at intermediate temperatures, and again these temperatures can formally be shown to correspond to the theta temperatures for these systems. This identification is useful in light of the recent results of Hofmann et al.[15] They propose that unfolded ensembles under native conditions are equivalent to those accessed by

proteins at theta temperatures. If this proposed equivalence holds up to scrutiny, then the analysis of ensembles generated for temperatures where $\Phi_T \rightarrow 0$ should lead to insights regarding unfolded states sampled under folding conditions.

The intrinsically stiff P_{56} system shows a gradual transition of Φ_T from a high value to a smaller value. As shown previously[49, 66], the increase in conformational heterogeneity arises mainly due to the increased frequency of generating bends and kinks within polyproline and this is partly due to the increased frequency of sampling *cis* peptide bonds at higher temperatures. Overall, however, the energy scales that encode chain stiffness in polyproline cannot be overcome by increasing temperature, and the transitions of both s^2 and Φ_T reflect this feature.

The lack of change in Φ_T with temperature for R_{56} is consistent with the maintenance of chain expansion over the entire temperature range as shown in Figure 6.1. Previous studies have shown that this system adopts expanded conformations whereby the degree of chain expansion exceeds that of self-avoiding random walks[47]. This expansion results from a combination of long-range electrostatic repulsions and favorable solvation of charged sidechains that together give rise to correlated fluctuations and the presence of multiple rod-like segments (stretches of polyproline II helices) within the chain[67]. In this scenario, chain compaction and increased conformational heterogeneity can be realized by screening the electrostatic repulsions in the presence of high concentrations of salt as was shown previously[47].

The preceding analysis clearly shows that it is insufficient to use Φ_T or s^2 alone to characterize the degree and nature of conformational heterogeneity. In systems such as NTL9 and GB1 the joint use of s^2 and Φ_T highlights the positive coupling between increased conformational heterogeneity and chain expansion. This is consistent with energy landscape

theories that predict strongly funneled landscapes for sequences that fold into well-defined ensembles of self-similar conformations[54]. Conversely, polyglutamine and polyarginine show limiting behaviors. For polyglutamine, joint use of s^2 and Φ_T shows that the globule-to-coil transition observed as temperature increases is consistent with ensembles switching from one class of heterogeneity to another. In the parlance of energy landscape theory, temperature modulates the ruggedness and free energy landscapes. For polyarginine, analysis of the temperature dependence of Φ_T alone seems confounding because it suggests a highly ordered system. However, when combined with the temperature dependence of s^2, we obtain a clearer inference regarding the ensembles, which highlights the contributions from electrostatic repulsions and the favorable solvation of charges sidechains that give rise to correlated fluctuations in highly charged systems.

6.3.4 Application of Φ_T to assess conformational heterogeneity in IDPs with different secondary structure propensities

Basic region leucine zippers (bZIPs) are modular transcription factors that play key roles in eukaryotic gene regulation[68]. The basic regions of bZIPs (bZIP-bRs) adopt regular α-helical conformations when bound to DNA[69]. Bioinformatics predictions and spectroscopic studies suggest that unbound, monomeric bZIP-bRs are uniformly disordered as autonomous units[70, 71]. This assumption was recently tested through quantitative characterization of the conformational preferences of fifteen different bZIP-bRs[60]. These were found to have quantifiable preferences for α-helical conformations in their unbound, monomeric forms. This helicity varies from one bZIP-bR to another despite significant sequence similarity of the DNA binding motifs (DBMs). Analysis of the determinants of helicity revealed that intramolecular interactions between DBMs

and 8-residue segments directly N-terminal to DBMs are the primary modulators of bZIP-bR helicities. The accuracy of this inference was tested in designed chimeras of bZIP-bRs that have either increased or decreased overall helicities. For a given sequence, the helical propensity $f_{\alpha}^{(T)}$ at temperature *T* was calculated using the formula in Equation (1):

$$f_{\alpha}^{(T)} = \frac{\sum_{i=1}^{N} \left\langle p_i^{(\alpha)} \right\rangle_T}{N}$$

$$\text{where } \left\langle p_i^{(\alpha)} \right\rangle_T = \left(\frac{\sum_{k=1}^{n_{\text{conf.}}^{(T)}} \Theta_k^i}{n_{\text{conf.}}^{(T)}} \right) \tag{1}$$

$$\text{and } \Theta_k^i = \begin{cases} 1, & \text{if residue } i \text{ is part of a helical segment in conformation } k \\ 0, & \text{otherwise} \end{cases}$$

In Equation (1), *N* denotes the number of residues in a bZIP-bR sequence, $\left\langle p_{\alpha}^{(i)} \right\rangle_i$ is the ensemble-averaged percent probability of finding residue *i* as part of a helical segment, $n_{\text{conf.}}^{(\text{T})}$ denotes the number of conformations used for calculating ensemble averages at temperature *T*, and Θ_k^i is a discrete Heaviside function that determines if residue *i* is part of an α-helical segment in conformation *k*. A α-helical segment was identified as a stretch that has *at least* seven consecutive residues that carry a DSSP (Define Secondary Structure of Proteins)[72] designation of "H", which implies that these residues are part of a regular, hydrogen-bonded α-helix. Panel A in Figure 6.6 shows a plot of Φ_T against the calculated helicity for seventeen bZIP-bRs that includes thirteen naturally occurring bZIP-bRs and four designed chimeric sequences. The results are shown for *T*=298 K.

It is intuitive to expect a strong positive correlation between an increase in Φ_T and an increase in helical propensity, although it is not clear that there should be a linear dependence of Φ_T on helical propensity or vice versa. In the interest of simplicity, we quantified the linear correlation between Φ_T and helical propensity using the Pearson product moment correlation coefficient. We find a value of $r = 5 \times 10^{-4}$ when we use the Φ_T and $f_\alpha^{(T)}$ values for all of the sequences listed in panel (A) of Figure 6.6. However, the situation improves to $r = 0.71$ when we restrict the correlation analysis to eleven out of the seventeen sequences. The outliers are as follows: The bZIP-bR of gcn4 shows high helicity (≈ 0.6) and low Φ_T (< 0.3). Conversely, the bZIP-bRs of fra1, fos, and opaque2 as well as the two chimeric bR sequences fos-gcn4, and fos-cys3 have low helicities (< 0.25) and higher values of Φ_T (> 0.4). We reasoned that the presence of significant outliers in the correlation analysis comes about because helicity is a probe of local structural propensities. Accordingly, it should follow that the quality of the correlation analysis improves, without discarding outliers, when one considers measures of overall conformational fluctuations. We therefore analyzed the correlation between Φ_T and the variance of the R_g distributions *i.e.*, $\sigma^2(R_g)$ for each of the seventeen sequences. Here, we expect a negative correlation between Φ_T and $\sigma^2(R_g)$ because increased conformational heterogeneity should lead to larger fluctuations in chain size and densities. Panel (B) in Figure 6.6 quantifies this negative correlation, which has a Pearson correlation coefficient of $r = -0.63$. This analysis clearly demonstrates the weaknesses inherent to using locally averaged measures of structure because it masks the degree of conformational heterogeneity that can be accommodated within an ensemble despite quantifiable secondary structure contents. In previous work Das et al.[60] used *de novo* sequence design to modulate intrinsic helicities of bZIP-bRs. Inasmuch as this effort was geared toward modulating the bias toward or away from α-helical conformations adopted by bZIP-bRs

in their bound states, the current analysis highlights the need to assess the degree of order / disorder in the ensemble by combining the calculation of helical propensities with an assessment of Φ_T. This in turn provides information about contributions to heterogeneity without focusing on a specific conformational preference.

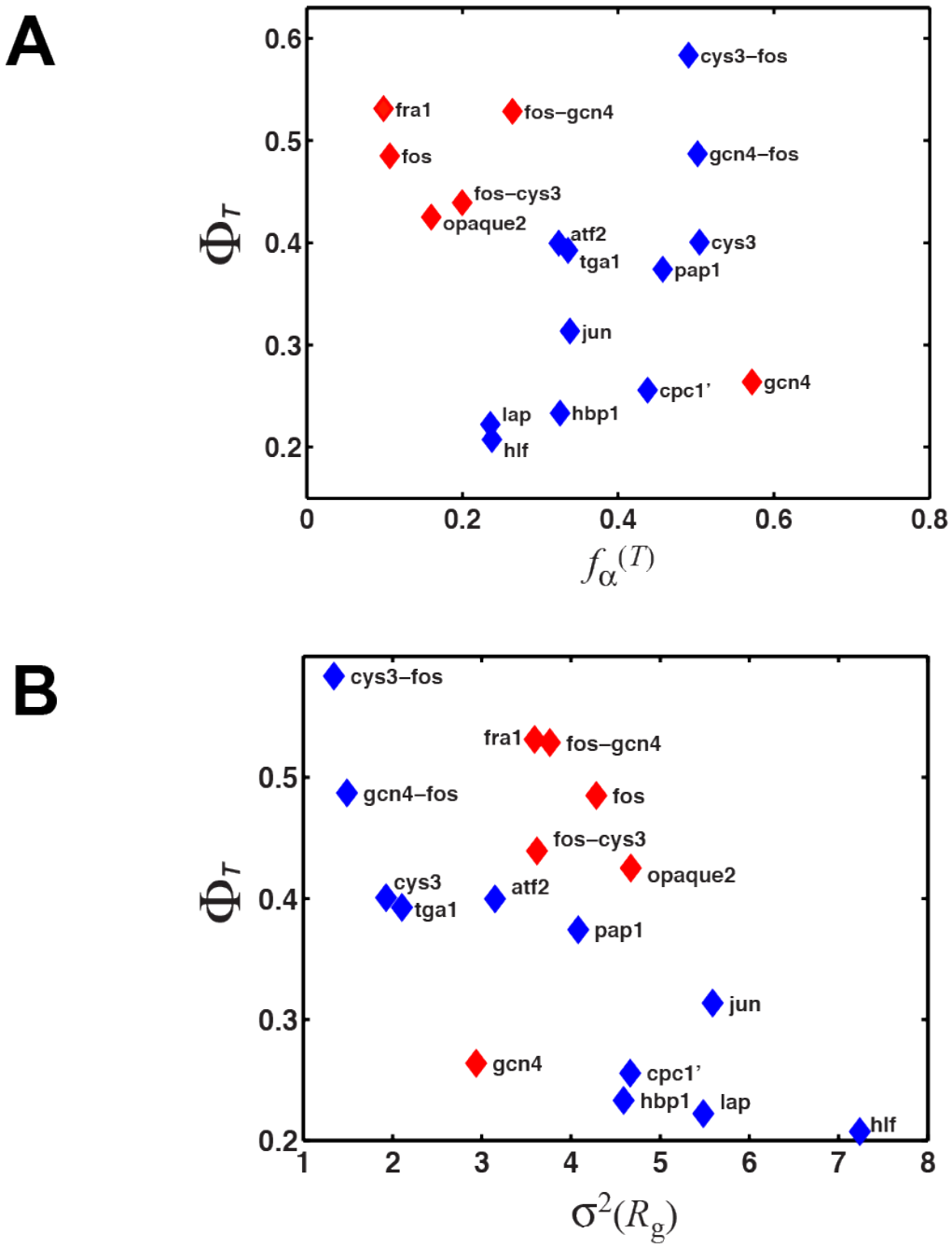


Figure 6.6: Correlations between Φ_T and measures of local structure versus measures of density fluctuations. Panel (A) plots Φ_T against $f_\alpha^{(T)}$ for $T = 298$ K. The latter is calculated using the DSSP algorithm as described in the main text. The results are shown for seventeen naturally occurring and designed sequences. Results for sequences shown in red cause significant

deviation from linear correlation between Φ_T and $f_\alpha^{(T)}$. The sequence identities for the bZIP-bRs and chimeras are shown adjacent to each of the symbols. Panel (B) plots Φ_T against $\sigma^2(R_g)$ for each of the seventeen bZIP-bRs. The color-coding of symbols is identical that of panel (A).

6.4 Discussion

We have introduced a parameter Φ_T that when combined with measures such as s^2 and quantification of local secondary structure preferences helps provide a clearer quantitative assessment of the degree of conformational heterogeneity in ensembles of proteins. We propose that this parameter will prove useful in comparative assessments of conformational heterogeneity of conformational ensembles generated for a single system at different temperatures and solution conditions as well as for different systems under similar conditions. To calculate Φ_T, we relied on three distinct choices namely, (i) the use of conformational vectors where the elements are inter-residue distances extracted from a specific conformation; (ii) the use of the distribution of pairwise projections of these vectors to calculate the degree of intra-ensemble dissimilarity; and (iii) the use of the FRC model to calibrate the degree of heterogeneity. We discussed the advantages of choice (iii) in the main text. Choices (i) and (ii) lead to an assessment of intra-ensemble conformational dissimilarity. Although these choices are distinct, they are not inherently superior to other methods proposed in the literature. For example, we could have calculated all unique pairwise superpositions of conformations based on least squares optimization and used the resultant distribution of root mean squared deviations (RMSDs)[73] as measures of dissimilarities. The computational expense of these calculations increases substantially with increased number of conformations in an ensemble. One could also use the method of projections to compare conformational vectors comprised of backbone dihedral angles

as elements[74]. We do not find any intrinsic advantages with using dihedral angle based conformational vectors and this could be used interchangeably with inter-residue distance based conformational vectors. Recent efforts have focused on the use of the number of inter-residue contacts q.[75] Each conformation within the ensemble is annotated by its q-number and the distributions of q-numbers *viz*., $P(q)$ are analyzed to compare different ensembles to each other. This method, which is analogous to methods used in spin glass theories[76], can be used in conjunction with Φ_T. It should be noted that the annotation of conformations by q-numbers requires the imposition of an *ad hoc* criterion for defining contacts, which causes an inherent loss of information. This is in contrast to the conformational vectors $\mathbf{V}_c$ used in this work.

6.4.1 Practical uses for Φ_T

The calculation of Φ_T is designed with two practical purposes in mind. As noted in the introduction, it is important to have measures of conformational heterogeneity that complement the assessments of ensembles that are obtained by quantification of densities, their fluctuations, and variances in energies. As useful as these parameters are, they can be incomplete as was demonstrated for systems that undergo coil-to-globule transitions without necessarily acquiring an ensemble of self-similar compact structures. In order to understand the mechanisms of coupled folding and binding of IDPs it would be useful to be able to modulate the degree of disorder in the unbound ensemble using *de novo* sequence design. The parameter Φ_T helps in this regard because it provides a direct measure of conformational heterogeneity and can be used to guide sequence design in a way that heterogeneity is either decreased (increased Φ_T) or increased (decreased Φ_T).

6.5 Acknowledgments

This work was supported by grants from the National Institutes of Health (5RO1NS056114) and the National Science Foundation (MCB-1121867).

6.6 Conclusions

Conformational heterogeneity is a defining characteristic of intrinsically disordered proteins (IDPs) and denatured states of proteins. Inferences regarding globule versus coil formation can be drawn from analysis of polymeric properties such as average size, shape, and density fluctuations. Here we introduce a new parameter to quantify the degree of conformational heterogeneity within an ensemble and complement polymeric descriptors of densities and density fluctuations. Each conformation in an ensemble is converted into an N-dimensional conformational vector where the elements are inter-residue distances and N is the number of unique distances. Similarity between pairs of conformations is quantified using the projection between the corresponding conformational vectors. An ensemble of conformations yields a distribution of pairwise projections, which are converted into a distribution of pairwise conformational dissimilarities. The first moment of this dissimilarity distribution is normalized against the first moment of the distribution obtained by comparing conformations from the ensemble of interest to conformations drawn from a Flory random coil model. The latter sets an upper bound on conformational heterogeneity thus ensuring that the proposed measure for intra-ensemble heterogeneity is properly calibrated and can be used to compare ensembles for different sequences and across different temperatures. Application of the new parameter to ensembles obtained as a function of temperature for different archetypal systems reveals a spectrum of transitions including order-to-disorder transitions and disorder-to-disorder

transitions. The latter occurs despite clear evidence for reversible coil-to-globule transitions and is realized for specific archetypal IDPs including sequences such as polar tracts. The new measure of conformational heterogeneity will be useful in quantitative studies of coupled folding and binding of IDPs and in *de novo* sequence design efforts that are geared toward controlling the degree of heterogeneity in unbound forms of IDPs.

6.7 References

1. Wolynes, P. G., Eaton, W. A., and Fersht, A. R. (2012) Chemical physics of protein folding, *Proceedings of the National Academy of Sciences of the United States of America 109*, 17770-17771.
2. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., and Obradovic, Z. (2001) Intrinsically disordered protein, *J Mol Graph Model 19*, 26-59.
3. Dyson, H. J., and Wright, P. E. (2002) Coupling of folding and binding for unstructured proteins, *Current Opinion in Structural Biology 12*, 54-60.
4. Halfmann, R., Alberti, S., Krishnan, R., Lyle, N., O'Donnell, C. W., King, O. D., Berger, B., Pappu, R. V., and Lindquist, S. (2011) Opposing Effects of Glutamine and Asparagine Govern Prion Formation by Intrinsically Disordered Proteins, *Molecular Cell 43*, 72-84.
5. Eliezer, D. (2009) Biophysical characterization of intrinsically disordered proteins, *Current Opinion In Structural Biology 19*, 23-30.

6. Sosnick, T. R., and Barrick, D. (2011) The folding of single domain proteins - have we reached a consensus?, *Curr. Opin. Struct. Biol. 21*, 12-24.

7. Vendruscolo, M. (2007) Determination of conformationally heterogeneous states of proteins, *Current Opinion In Structural Biology 17*, 15-20.

8. Mao, A. H., Lyle, N., and Pappu, R. V. (2013) Describing sequence-ensemble relationships for intrinsically disordered proteins, *Biochemical Journal 449*, 307-318.

9. Anil, B., Li, Y., Cho, J. H., and Raleigh, D. P. (2006) The unfolded state of NTL9 is compact in the absence of denaturant, *Biochemistry 45*, 10110-10116.

10. Meng, W., Luan, B., Lyle, N., Pappu, R. V., and Raleigh, D. P. (2013) The Denatured State Ensemble Contains Significant Local and Long-Range Structure under Native Conditions: Analysis of the N-Terminal Domain of Ribosomal Protein L9, *Biochemistry 52*, 2662-2671.

11. Voelz, V. A., Jaeger, M., Yao, S., Chen, Y., Zhu, L., Waldauer, S. A., Bowman, G. R., Friedrichs, M., Bakajin, O., Lapidus, L. J., Weiss, S., and Pande, V. S. (2012) Slow Unfolded-State Structuring in Acyl-CoA Binding Protein Folding Revealed by Simulation and Experiment, *Journal of the American Chemical Society 134*, 12565-12577.

12. Hoffmann, A., Nettels, D., Clark, J., Borgia, A., Radford, S. E., Clarke, J., and Schuler, B. (2011) Quantifying heterogeneity and conformational dynamics from single molecule FRET of diffusing molecules: recurrence analysis of single particles (RASP), *Physical Chemistry Chemical Physics 13*, 1857-1871.

13. Lapidus, L. J. (2013) Exploring the top of the protein folding funnel by experiment, *Current Opinion in Structural Biology 23*, 30-35.

14. Meng, W., Lyle, N., Luan, B., Raleigh, D. P., and Pappu, R. V. (2013) Experiments and simulations show how long-range contacts can form in expanded unfolded proteins with negligible secondary structure, *Proceedings of the National Academy of Sciences of the United States of America 110*, 2123-2128.

15. Hofmann, H., Soranno, A., Borgia, A., Gast, K., Nettels, D., and Schuler, B. (2012) Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy, *Proc. Natl. Acad. Sci. U. S. A. 109*, 16155-16160.

16. Kohn, J. E., I.S. Millett, J. Jacob, B. Zagrovic, T.M. Dillon, N. Cingel, R.S. Dothager, S. Seifert, P. Thiyagarajan, T.R. Sosnick, M.Z. Hasan, V.S. Pande, I. Ruzcinski, S. Doniach, and Plaxco, K. W. (2004) Random-coil behavior and the dimensions of chemically unfolded proteins, *Proceedings of the National Academy of Sciences of the United States of America 101*, 12491-12496.

17. Zhou, H.-X., Rivas, G., and Minton, A. P. (2008) Macromolecular crowding and confinement: Biochemical, biophysical, and potential physiological consequences, *Annual Review of Biophysics 37*, 375-397.

18. Elcock, A. H. (2010) Models of macromolecular crowding effects and the need for quantitative comparisons with experiment, *Curr. Opin. Struct. Biol. 20*, 196-206.

19. Jahn, T. R., and Radford, S. E. (2008) Folding versus aggregation: Polypeptide conformations on competing pathways, *Archives of Biochemistry and Biophysics 469*, 100-117.

20. Lapidus, L. J. (2013) Understanding protein aggregation from the view of monomer dynamics, *Molecular Biosystems 9*, 29-35.

21. Dedmon, M. M., Lindorff-Larsen, K., Christodoulou, J., Vendruscolo, M., and Dobson, C. M. (2005) Mapping long-range interactions in alpha-synuclein using spin-label NMR and ensemble molecular dynamics simulations, *Journal of the American Chemical Society 127*, 476-477.

22. Soranno, A., Buchli, B., Nettels, D., Cheng, R. R., Mueller-Spaeth, S., Pfeil, S. H., Hoffmann, A., Lipman, E. A., Makarov, D. E., and Schuler, B. (2012) Quantifying internal friction in unfolded and intrinsically disordered proteins with single-molecule spectroscopy, *Proceedings of the National Academy of Sciences of the United States of America 109*, 17800-17806.

23. Dyson, H. J., and Wright, P. E. (2005) Intrinsically unstructured proteins and their functions, *Nat. Rev. Mol. Cell Biol. 6*, 197-208.

24. Pancsa, R., and Tompa, P. (2012) Structural Disorder in Eukaryotes, *Plos One 7*.

25. Tompa, P., and Fuxreiter, M. (2008) Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions, *Trends In Biochemical Sciences 33*, 2-8.

26. Mittag, T., Kay, L. E., and Forman-Kay, J. D. (2010) Protein dynamics and conformational disorder in molecular recognition, *Journal of Molecular Recognition 23*, 105-116.

27. Grosberg, A. Y., and Kuznetsov, D. V. (1992) Quantitative Theory Of The Globule-To-Coil Transition .1. Link Density Distribution In A Globule And Its Radius Of Gyration, *Macromolecules 25*, 1970-1979.

28. Sanchez, I. C. (1979) Phase transition behavior of the isolated polymer chain, *Macromolecules 12*, 980-988.

29. Gianni, S., Guydosh, N. R., Khan, F., Caldas, T. D., Mayor, U., White, G. W. N., DeMarco, M. L., Daggett, V., and Fersht, A. R. (2003) Unifying features in protein-folding mechanisms, *Proceedings of the National Academy of Sciences of the United States of America 100*, 13286-13291.

30. Sherman, E., and Haran, G. (2006) Coil-globule transition in the denatured state of a small protein, *Proceedings of the National Academy of Sciences of the United States of America 103*, 11539-11543.

31. Ziv, G., Thirumalai, D., and Haran, G. (2009) Collapse transition in proteins, *Physical Chemistry Chemical Physics 11*, 83-93.

32. Shea, J. E., and Brooks, C. L. (2001) From folding theories to folding proteins: A review and assessment of simulation studies of protein folding and unfolding, *Annual Review of Physical Chemistry 52*, 499-535.

33. O'Brien, E. P., Brooks, B. R., and Thirumalai, D. (2009) Molecular Origin of Constant m-Values, Denatured State Collapse, and Residue-Dependent Transition Midpoints in Globular Proteins, *Biochemistry 48*, 3743-3754.

34. Nettels, D., Gopich, I. V., Hoffmann, A., and Schuler, B. (2007) Ultrafast dynamics of protein collapse from single-molecule photon statistics, *Proceedings of the National Academy of Sciences of the United States of America 104*, 2655-2660.

35. Sinha, K. K., and Udgaonkar, J. B. (2005) Dependence of the size of the initially collapsed form during the refolding of barstar on denaturant concentration: evidence for a continuous transition, *J Mol Biol 353*, 704-718.

36. Chou, J. J., and Shakhnovich, E. I. (1999) A study on local-global cooperativity in protein collapse, *Journal of Physical Chemistry B 103*, 2535-2542.

37. Udgaonkar, J. B. (2013) Polypeptide chain collapse and protein folding, *Archives of Biochemistry and Biophysics 531*, 24-33.

38. Crick, S. L., Jayaraman, M., Frieden, C., Wetzel, R., and Pappu, R. V. (2006) Fluorescence correlation spectroscopy shows that monomeric polyglutamine molecules form collapsed structures in aqueous solutions, *Proceedings Of The National Academy Of Sciences Of The United States Of America 103*, 16764-16769.

39. Mukhopadhyay, S., Krishnan, R., Lemke, E. A., Lindquist, S., and Deniz, A. A. (2007) A natively unfolded yeast prion monomer adopts an ensemble of collapsed and rapidly fluctuating structures, *Proceedings of the National Academy of Sciences of the United States of America 104*, 2649-2654.

40. Teufel, D. P., Johnson, C. M., Lum, J. K., and Neuweiler, H. (2011) Backbone-Driven Collapse in Unfolded Protein Chains, *Journal of Molecular Biology 409*, 250-262.

41. Marsh, J. A., and Forman-Kay, J. D. (2010) Sequence Determinants of Compaction in Intrinsically Disordered Proteins, *Biophysical Journal 98*, 2383-2390.

42. Jain, N., Bhattacharya, M., and Mukhopadhyay, S. (2011) Chain Collapse of an Amyloidogenic Intrinsically Disordered Protein, *Biophys. J. 101*, 1720-1729.

43. Brocca, S., Testa, L., Sobott, F., Samalikova, M., Natalello, A., Papaleo, E., Lotti, M., De Gioia, L., Doglia, S. M., Alberghina, L., and Grandori, R. (2011) Compaction Properties of an Intrinsically Disordered Protein: Sic1 and Its Kinase-Inhibitor Domain, *Biophysical Journal 100*, 2243-2252.

44. Vaiana, S. M., Best, R. B., Yau, W.-M., Eaton, W. A., and Hofrichter, J. (2009) Evidence for a Partially Structured State of the Amylin Monomer, *Biophysical Journal 97*, 2948-2957.

45. Vitalis, A., Lyle, N., and Pappu, R. V. (2009) Thermodynamics of beta-Sheet Formation in Polyglutamine, *Biophys. J. 97*, 303-311.

46. Muller-Spath, S., Soranno, A., Hirschfeld, V., Hofmann, H., Ruegger, S., Reymond, L., Nettels, D., and Schuler, B. (2010) Charge interactions can dominate the dimensions of intrinsically disordered proteins, *Proceedings of the National Academy of Sciences of the United States of America 107*, 14609-14614.

47. Mao, A. H., Crick, S. L., Vitalis, A., Chicoine, C. L., and Pappu, R. V. (2010) Net charge per residue modulates conformational ensembles of intrinsically disordered proteins, *Proceedings of the National Academy of Sciences of the United States of America 107*, 8183-8188.

48. Tooke, L., Duitch, L., Measey, T. J., and Schweitzer-Stenner, R. (2010) Kinetics of the Self-Aggregation and Film Formation of Poly-L-Proline at High Temperatures Explored by Circular Dichroism Spectroscopy, *Biopolymers 93*, 451-457.

49. Radhakrishnan, A., Vitalis, A., Mao, A. H., Steffen, A. T., and Pappu, R. V. (2012) Improved atomistic Monte Carlo simulations demonstrate that poly-L-proline adopts heterogeneous ensembles of conformations of semi-rigid segments interrupted by kinks, *Journal of Physical Chemistry B 116*, 6862-6871.

50. Vitalis, A., Wang, X., and Pappu, R. V. (2008) Atomistic simulations of the effects of polyglutamine chain length and solvent quality on conformational equilibria and spontaneous homodimerization, *J. Mol. Biol. 384*, 279-297.

51. Chen, S., Berthelier, V., Yang, W., and Wetzel, R. (2001) Polyglutamine aggregation behavior in vitro supports a recruitment mechanism of cytotoxicity, *Journal Of Molecular Biology 311*, 173-182.

52. Bryngelson, J. D., Onuchic, J. N., Socci, N. D., and Wolynes, P. G. (1995) Funnels, Pathways, And The Energy Landscape Of Protein-Folding - A Synthesis, *Proteins-Structure Function And Genetics 21*, 167-195.

53. Onuchic, J. N., LutheySchulten, Z., and Wolynes, P. G. (1997) Theory of protein folding: The energy landscape perspective, *Annual Review of Physical Chemistry 48*, 545-600.

54. Onuchic, J. N., and Wolynes, P. G. (2004) Theory of protein folding, *Current Opinion in Structural Biology 14*, 70-75.

55. Vitalis, A., Wang, X., and Pappu, R. V. (2007) Quantitative characterization of intrinsic disorder in polyglutamine: Insights from analysis based on polymer theories, *Biophys. J. 93*, 1923-1937.

56. Papoian, G. A. (2008) Proteins with weakly funneled energy landscapes challenge the classical structure-function paradigm, *Proceedings of the National Academy of Sciences of the United States of America 105*, 14237-14238.

57. Potoyan, D. A., and Papoian, G. A. (2011) Energy Landscape Analyses of Disordered Histone Tails Reveal Special Organization of Their Conformational Dynamics, *Journal of the American Chemical Society 133*, 7405-7415.

58. Camacho, C. J., and Thirumalai, D. (1993) Minimum energy compact structures of random sequences of heteropolymers, *Physical Review Letters 71*, 2505-2508.

59. Chan, H. S., and Dill, K. A. (1993) Energy Landscapes And The Collapse Dynamics Of Homopolymers, *Journal of Chemical Physics 99*, 2116-2127.

60. Das, R. K., Crick, S. L., and Pappu, R. V. (2012) N-terminal segments modulate the alpha-helical propensities of the intrinsically disordered basic regions of bZIP proteins, *J. Mol. Biol. 416*, 287-299.

61. N. Metropolis, A. R., M. Rosenbluth, A. Teller, and E. Teller. (1953) Equation of state calculations by fast computing machines, *Journal of Chemical Physics 21*, 1087-1092.

62. Vitalis, A., and Pappu, R. V. (2009) ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions, *J. Comput. Chem. 30*, 673-699.

63. Kaminski, G. A., Friesner, R. A., Tirado-Rives, J., and Jorgensen, W. L. (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides, *Journal Of Physical Chemistry B 105*, 6474-6487.

64. Engh, R. A., and Huber, R. (1991) Accurate bond and angle parameters for X-ray protein structure refinement, *Acta Cryst. A47*, 392-400.

65. Flory, P. J. (1969) *Statistical Mechanics of Chain Molecules*, Oxford University Press, New York.

66. Best, R. B., Merchant, K. A., Gopich, I. V., Schuler, B., Bax, A., and Eaton, W. A. (2007) Effect of flexibility and cis residues in single-molecule FRET studies of polyproline, *Proceedings of the National Academy of Sciences of the United States of America 104*, 18964-18969.

67. Ha, B. Y., and Thirumalai, D. (1992) Conformations of a polyelectrolyte chain, *Physical Review A 46*, R3012-R3015.

68. Amoutzias, G. D., Veron, A. S., Weiner, J., Robinson-Rechavi, M., Bornberg-Bauer, E., Oliver, S. G., and Robertson, D. L. (2007) One billion years of bZIP transcription factor evolution: Conservation and change in dimerization and DNA-binding site specificity, *Molecular Biology and Evolution 24*, 827-835.

69. Ellenberger, T. E., Brandl, C. J., Struhl, K., and Harrison, S. C. (1992) The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted alpha-helices - crystal-structure of the protein-DNA complex, *Cell 71*, 1223-1237.

70. Liu, J. G., Perumal, N. B., Oldfield, C. J., Su, E. W., Uversky, V. N., and Dunker, A. K. (2006) Intrinsic disorder in transcription factors, *Biochemistry 45*, 6873-6888.

71. Oneil, K. T., Hoess, R. H., and Degrado, W. F. (1990) DESIGN OF DNA-BINDING PEPTIDES BASED ON THE LEUCINE ZIPPER MOTIF, *Science 249*, 774-778.

72. Kabsch, W., and Sander, C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features., *Biopolymers 22*, 2577-2637.

73. Vriend, G., and Sander, C. (1991) DETECTION OF COMMON 3-DIMENSIONAL SUBSTRUCTURES IN PROTEINS, *Proteins-Structure Function and Bioinformatics 11*, 52-58.

74. Mu, Y. G., Nguyen, P. H., and Stock, G. (2005) Energy landscape of a small peptide revealed by dihedral angle principal component analysis, *Proteins-Structure Function and Bioinformatics 58*, 45-52.

75. Potoyan, D. A., and Papoian, G. A. (2012) Regulation of the H4 tail binding and folding landscapes via Lys-16 acetylation, *Proceedings of the National Academy of Sciences of the United States of America 109*, 17857-17862.

76. Parisi, G. (1983) ORDER PARAMETER FOR SPIN-GLASSES, *Physical Review Letters 50*, 1946-1948.

Chapter 7

Summary of contributions and future directions

This dissertation contributes to our understanding of the nature and role of conformational heterogeneity in proteins. Such insights are relevant to understanding protein folding, self-assembly, in addition to regulatory and disease processes. The main contributions can be grouped into into three categories. This first affords a complete characterization of the denatured and non-native states of folded or intrinsically disordered proteins. The seconds describes how conformational heterogeneity facilitates protein aggregation and self-assembly. Lastly, we provide methods to quantify the degree and nature of conformational heterogeneity for protein ensembles and describe novel analysis techniques that bridge simulation results to biophysical measurements. Together, these improvements provide design handles on conformational heterogeneity. Going forward, this allows new areas of investigation focused on the modulation of heterogeneity.

The salient findings described in this work have broad applicability to different areas of biophysics. Chapter 2 describes the computational generation of all-atom resolution denatured state ensembles that provide detailed information and testable hypotheses regarding denatured state interactions. The uniqueness of our approach is that our simulation model does not depend on input from experimental measurements to guide the calculations. These ensembles are not post processed according to ad-hoc filtering criteria and they are generated with no biases other than the underlying ABSINTH Hamiltonian. This has previously proven robust in generating protein ensembles consistent with measured experimental quantities. In our analysis of the NTL9 denatured

state, we find that clusters of hydrophobic residues are responsible for long-range interactions in the denatured state and we posit these interactions prevent deleterious interactions that may result in misfolding or aggregation. It is likely these interactions are present in the denatured state ensembles of other folded proteins as well, which suggests they play a role in modulating the barriers, rates, and pathways involved in protein folding. Another interesting and hopefully not overlooked point is that *satisfaction of good solvent chain scaling behavior is compatible with long-range attractive interactions* and this explains observed deviations from predictions made by excluded volume or worm like chain models of denatured states. Identification of the nature of these interactions allows one to propose sequence mutants that alter the stability of the denatured state. Since this is thermodynamically linked to the stability of the folded state, we have a new tool that allows tuning of the stability of folded proteins. This circumvents analysis of native state interactions alone, which could be a limiting prospect in the absence of structural data. Chapter 3 shows that intramolecular disorder-to-order transitions are not required to for intermolecular associations and aggregation of polyglutamine and polyglutamine-rich proteins. The lag-times associated with nucleated conformational conversions to β-sheet rich fibrils are due to the slow dynamics of reptating chains in the middle of large clusters of molecules. Chapter 4 titrates the role of sequence context in modulating conformational heterogeneity and how this directs the self-assembly of glutamine and asparagine rich protein aggregates. Molecules with an increased fraction of glutamine residues are enriched in heterogeneity. Such results are consistent with the destabilizing effect imparted by polyglutamine expansions on their host proteins as they have a locally denaturing effect. This increased heterogeneity for

glutamine-rich constructs drives the formation of smaller oligomers that are coincident with an increase in toxicity to yeast cultures. The shorter asparagine sidechain reduces conformational heterogeneity and promotes β-hairpin turn formation. This results in ordered self-templating amyloid aggregates that were benign. This provides further support for the *toxic oligomer* hypothesis, whereby the promiscuity of interactions seen with polyglutamine-rich soluble oligomers allows for deleterious interactions with unintended partners. Polyglutamine stretches play a role in several transcriptional regulatory pathways and their disruption is an obvious route to toxicity. Chapter 5 shows that conformational heterogeneity and therefore intermolecular associations can be modified by naturally occurring flanking sequences that flank polyglutamine expansions associated with several neurodegenerative disorders. This provides a therapeutic route to prevent intermolecular associations that could be repurposed to slow the onset of neuronal damage. Another prospect from this chapter is that it provides a framework that makes use of all atom simulation data from of monomers, dimers, and (at a later date) trimers, to develop coarse-grained pair potentials for aggregating molecules. This allows one to account for flanking-sequence effects and molecular deformations that determine the size and shape preferences of aggregation intermediates and their role in shaping the aggregation mechanism. Chapter 6 provides a numerical measure for conformational heterogeneity in a protein conformational ensemble that we refer to as Φ. This measure is able to distinguish between systems that strongly couple their unfolding-folding transitions with coil-to-globule transitions and those systems that undergo coil-to-globule transitions with no evidence of acquiring an ensemble of self-similar conformations upon collapse.