5-24-2009

Probing the Early Stages of Polyglutamine Aggregation with Computational Methods

Andreas Vitalis
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences

Molecular Biophysics

Dissertation Examination Committee:
Rohit V. Pappu, Chair
Nathan A. Baker
Anders E. Carlsson
Roberto Galletto
Lev D. Gelb
Timothy M. Lohman

PROBING THE EARLY STAGES OF POLYGLUTAMINE AGGREGATION WITH COMPUTATIONAL METHODS

by

Andreas Vitalis

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 2009

Saint Louis, Missouri

ABSTRACT OF THE DISSERTATION

Probing the Early Stages of Polyglutamine Aggregation

with Computational Methods

Exonic CAG repeat diseases are a class of neurodegenerative age-of-onset diseases caused by an unstable trinucleotide expansion in a coding region of a gene. The most prominent example is Huntington's disease (HD) whose symptoms are characterized by loss of motor control and cognitive deficits. For all nine of the known CAG repeat diseases, pathology is ascribed to the mutant proteins which carry expanded stretches of glutamine residues (polyglutamine). The length of the polyglutamine segment is inversely correlated with the disease age-of-onset. Protein aggregates are routinely found in *postmortem* tissue samples of brains of HD patients. These findings suggest a prominent role for polyglutamine-mediated protein aggregation in disease pathogenesis.

Subsequent studies characterized the intracellular aggregates as amyloid-like. In amyloids, the polypeptide backbone predominantly adopts conformations in the β-basin of the Ramachandran map, *i.e.*, the aggregates have high net β-content. This has led to the hypothesis that β-rich conformers play a prominent role in mediating the aggregation process; specifically, it has been postulated that a β-rich form of polyglutamine acts as the monomeric nucleus from which fibrillar aggregates grow via a downhill elongation mechanism.

This thesis investigates the intrinsic properties of polyglutamine during early stages of aggregation. We employ computer simulations to obtain a qualitative picture of the process at an atomistic level. Our results suggest the following: soluble polyglutamine is intrinsically disordered and forms collapsed globules in aqueous solution. These globules associate readily and randomly to form disordered dimers. We identified no structural requirements for association to occur. The conversion of monomeric polyglutamine to a conformation high in β-content, *i.e.*, to a putative aggregation nucleus, is associated with a high free energy penalty. We detect no coupling between structure and associativity, but find a profound modulation of polyglutamine's intrinsic properties in the presence of wild-type flanking sequences.

From our results, we postulate a model where polyglutamine forms large soluble and disordered oligomers which undergo a rate-limiting conformational conversion to a fibrillar precipitate. We conclude that structure-based drug designs may not prove a viable strategy for interfering with the early stages of polyglutamine aggregation and hence with disease pathology.

ACKNOWLEDGMENTS

First of all, I would like to acknowledge my mentor, Rohit, who has unconditionally supported my scientific activities throughout our time together. His obsession with solving problems, with scientific rigor, and with interfacing our work with that of others has been most inspiring. However, his ability to remain a human being within the constraints of pre-tenure stress, political battles, and emotional turbulence is what truly offsets him in my mind. The warmth and appreciation, in particular in the early days, will always be fondly remembered.

Next, I would like to acknowledge fellow members of the lab of past and present. Hoang, Alan, Scott, Tim, Albert and Matt have helped create a stimulating and enjoyable environment. Xiaoling and Nick I am particularly indebted to for their willingness not only to share a workspace with me but to also collaborate in the truest sense of the word. The contributions of undergraduates and rotation students shall not be forgotten: Emma, Robert, and Jose in particular have earned my gratitude for being willing to work with me on projects not necessarily garnished with the potential for individual glory. Adam has been in a unique position as a technician and deserves acknowledgment for his commitment to the job. Finally, Alan, and more recently Nick as well, deserve particular mention for carrying their share of the administrative load – a duty which can not even be remotely expected to be fulfilled so willingly by students in pursuit of their own scientific goals.

Without funding predominantly from the NIH (grant 5R01-NS056114) but also from the NSF (grant MCB 0718924), this work would not have been possible and the missions of both agencies are duly acknowledged here. I would like to thank my thesis committee members for their willingness to follow my scientific pursuits over the years and for helpful suggestions during update meetings.

Times have not always been easy but I do wish to express my gratitude to the people I have met and who have made me laugh and think: Matt (and Katie), Albert, Hoang, Todd (and Diana), and Alan, and in particular Alyssa, Erin, Jeff, and Mike.

Most importantly, I have to and want to acknowledge those people who have made me live: my family, first and foremost, for their absolutely unwavering support: not only of my work and my life, but also of my decision to separate myself from them geographically so much and for such a long time. We never lost touch, and this has been essential for me personally. Enormous gratitude I owe to Miriam who had (and probably has) the unique talent to make me think, to make me question my own ways. I would not be who I am today were it not for the time we spent both together and apart. I am deeply grateful to Volker for being a true friend over the years. His wit and support have guided me through most difficult times. Lastly, I have found true happiness in my relationship with Noelle. I am deeply grateful to her for accepting me as the person I am today. The emotional support I have derived from my relationship with her and with the aforementioned human beings I could not describe accurately even if I tried to.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

Page

LIST OF TABLES

ABBREVIATIONS AND TERMS

(terms appear in lexicographical order)

ABSINTH: Self-assembly of biomolecules studied by a novel, implicit, and tunable Hamiltonian

AD: Alzheimer's disease

AMBER: Assisted model building with energy refinement

BAG1: B-cell leukemia/lymphoma 2-associated athanogene

β-secondary structure: Ensemble of possible secondary structure elements with canonical β-hydrogen bond registry and distinct population of the β-basin in Ramachandran space including β-sheets, β-hairpins and β-helices

CAMPARI: Computational analysis of macromolecular properties across resolutions and interfaces

CD: Circular dichroism

CHARMM: Chemistry at Harvard molecular mechanics

CHIP: C-terminus of Hsp70-interacting protein

CRABP: Cellular retinoic acid binding protein

DLS: Dynamic light scattering

DMFI: Direct mean-field interaction

DNA: Deoxyribonucleic acid

DRPLA: Dentatorubral pallidoluysian atrophy

DSSP: Define secondary structure of proteins

DSSP E-score: Fraction of residues engaging in hydrogen bonds with registry characteristic for β-secondary structure

EEF1: Effective energy function 1

EV: Excluded volume

FCS: Fluorescence correlation spectroscopy

FRET: Förster resonance energy transfer

GB: Generalized Born

GFP: Green fluorescent protein

GROMACS: Groningen machine for chemical simulations

GROMOS: Groningen molecular simulation computer program package

GST: Glutathione S-transferase

HD: Huntington's disease

HFIP: Hexafluoroisopropanol

HPLC: High-pressure liquid chromatography

Hsp: Heat shock protein

htt: Huntingtin

hydrophile: Hydrophilic polypeptide residue

hydrophobe: Hydrophobic polypeptide residue

HYPK: Huntingtin yeast two-hybrid protein K

ICA: Independent component analysis

IDP: Intrinsically disordered protein

LD: Langevin dynamics

LJ: Lennard-Jones

MC: Monte Carlo

MD: Molecular dynamics

MIM: Mendelian inheritance in man

MMC: Metropolis Monte Carlo

MRMD: Multiple replica molecular dynamics

NMR: Nuclear magnetic resonance

Nt17: The 17 residues N-terminal to the polyQ-expansion on exon1 of htt

OPLS: Optimized potential for liquid simulations

OPLS-AA: Optimized potential for liquid simulations – all-atom

OPLS-UA: Optimized potential for liquid simulations – united-atom

PB: Poisson-Boltzmann

PCA: Principal component analysis

PMF: Potential of mean force

polyglutamine: Polypeptides composed almost exclusively of glutamine

polyQ: See **polyglutamine**

PQCS: Protein quality control system

QBP1: polyQ binding peptide 1

Q_N: Acetyl-$(Gln)_N$-*N*-methylamide

REX: Replica exchange

R_g: Radius of gyration

RMSD: Root mean square deviation

RNA: Ribonucleic acid

SA: Surface area

SASA: Solvent-accessible surface area

SAV: Solvent-accessible volume

SAXS: Small-angle X-ray scattering

SBMA: Spinal and bulbar muscular atrophy

SCA: Spinocerebellar ataxia

SDS: Sodium dodecyl sulfate

SE: Standard error

SUMO: Small ubiquitin-like modifier

TFA: Trifluoroacetic acid

TFE: Trifluoroethanol

ThT: Thioflavin T

TI: Thermodynamic integration

TRiC: T-complex polypeptide 1 ring complex

WHAM: Weighted histogram analysis method

CHAPTER I. INTRODUCTION

I.1. Preamble

Ideally, a doctoral thesis is the coherent sequence of novel and reproducible research on a system of relevance to the field of study. It is placed in a broader thematic context and is expected to advance the field significantly. A thesis remains distinguished from a journal article by its focus on a connected set of questions rather than a select few. Answers to all of them would simply extend beyond the scope of a single article, both in terms of sheer quantity of data and in terms of the impact of the results to the overarching storyline.

The present thesis is presented toward the partial fulfillment of the degree requirements for a Ph.D. in Molecular Biophysics. The system of relevance we chose is – not surprisingly – a biological one studied at the molecular level. In the broadest sense, it is that of protein aggregation diseases. More specifically, we targeted CAG-repeat diseases,[1] *i.e.*, diseases in which pathological aggregation occurs due to the translation of an extended polypeptide stretch in the host protein. This stretch is composed entirely of glutamine residues due to the repeated CAG-codons on the disease gene being translated.[2] As is outlined in I.3, we asked and attempted to answer very specific questions about the basic physicochemical mechanism of polyglutamine aggregation employing computer simulations as our solitary tool.

It is very important to point out that we have attempted to remain deeply rooted in the knowledge provided by the vast body of both *in vitro* and *in vivo*

experimental work. Throughout this multi-year endeavor, we have constantly evaluated our numerical and theoretical approaches in this manner, since only then can we confidently satisfy the requirements of both producing relevant results and of advancing a specific field significantly. The second point requires some elaboration: Biology poses the unique challenge that it makes conceptualizations extremely difficult due to its innate complexity. Conversely, physicists routinely attempt to simplify the problem to a level where such conceptualizations become feasible. Undoubtedly, one may ultimately succeed in defining a framework and a resolution at which feasibility is obtained. However, to provide a significant advancement of a *specific* area of biology, it is not only necessary to find such a framework, but also to make sure that it still describes the *specific* system with qualitative accuracy.

This last point is best illustrated using an example: Consider the problem of protein folding, which is much narrower in biological complexity than that of CAG-repeat disease pathology. It has lent itself to extremely helpful conceptualizations from a biophysical point of view; it has, for example, spawned the entire field of protein energy landscape theory.[3] In terms of a generalized theoretical framework, simplistic models such as minimalist lattice models have been immensely helpful, since they have the power to elucidate underlying and unifying concepts.[4] They do, however, fail in answering specific questions about the folding mechanism of a “real” protein. Instead, they facilitate the formulation of hypotheses for said “real” protein, which can then be tested using a more accurate numerical model or – of course – experimental techniques.

The above example demonstrates the fragile juncture at which the field of biophysics is placed in 2009. The physicists in us are drawn to the conceptualizations and principles, to the unifying mechanisms and driving forces at work. Conversely, the biologists in us are drawn to the details of the specific system under study, toward the primary goal of understanding precisely that system and not necessarily anything else. We have tried to benefit from the implied synergy in understanding the molecular mechanism of the pathology of CAG-repeat diseases. Thinking about physical driving forces gave rise to hypotheses which we tested computationally. The results helped us understand and reinterpret experimental data reliant too much on qualitative speculation otherwise. They also helped experimental colleagues think about new experiments of utmost biological relevance.

However, this fundamentally interdisciplinary process remains fraught with difficulty. Primarily, there are communication barriers which need to be eliminated for biophysicists to be truly operating at the interface. Transfer of concepts and knowledge in either direction is severely hampered by simple language barriers. Prejudice against different “schools of thought” presents additional hurdles. Much like modern scientists in a competitive funding environment are asked to be salesmen and –women of their own talents, they are equally asked to be diplomats, translators and even mediators. It appears as if the ability to quickly “export” one’s own set of methodologies, tools, and knowledge into an unfamiliar problem setting represents one of the fundamentally important skills a 21st-century scientist must acquire to successfully operate at the interface of

disparate areas of research and – sometimes – disparate schools of thought. Without that ability, biophysics might remain a subdivision such as physical biology or biological physics for the time being.

The remainder of this thesis is organized as follows: in the subsequent parts of Chapter I we introduce the topic of polyglutamine expansion diseases and review the literature as of 2009. While it would be easier to motivate hypotheses and methods for each of the additional chapters (II to VI) given the state of the field *at the time the particular projects were started*, it would also impose a historical tone onto the thesis. Instead, we opt to provide a *current* overview which clearly places our results in a *current* context. Our results are presented in Chapters II to VI. The work in Chapter II to V has been published previously in peer-reviewed journals,[5-8] while the manuscript for the material presented in Chapter VI is being prepared for submission in the very near future.[9] The published articles are used directly in those chapters with minor modifications intended to preserve the flow of content and logic of this thesis. In particular, Chapters II to VI each have a preamble meant to place them in the broader context of the thesis, to report alternative approaches which were pursued but proved unfruitful, and to point out the contributions of co-authors whenever other researchers beyond my advisor and myself were involved. The introduction appearing in the published articles is often (at least partially) removed, since a unifying introduction is provided here in Chapter I. Finally, Chapter VII summarizes the relevance of our efforts and looks ahead to future projects which have been or will be spawned by this work.

With the exception of this particular sentence, I have chosen to use the "we"-form for the entirety of this thesis. The reason is simple: while it might appear desirable to always attempt to decompose the efforts of individuals, science is teamwork, and a laboratory like ours represents a thinking environment fueled by incessant communication, *i.e.*, it represents a fundamentally collaborative model of productivity. That said, we will try to make it as clear possible to delineate individual co-authors' contributions as stated above.

I.2. Polyglutamine Expansion Diseases

I.2.1. Overview

Trinucleotide repeat diseases[1] derive their name from their unifying genetic feature that a specific codon is repeated on a gene multiple times. The repeat is unstable and can expand leading to a much improved susceptibility of the host organism to exhibit a disease phenotype. This connection was not understood until the early 1990s when the causative genes for X-linked spinal and bulbar muscular atrophy (SBMA),[10] Huntington's disease (HD),[11] and spinocerebellar ataxia type 1 (SCA1)[12] were identified. Since then, further diseases have been characterized. Table 1.a summarizes those in which the trinucleotide repeat is exonic, *i.e.*, is actually translated to yield a mutant protein. In all those cases, the unstable codon is CAG, which results in expanded polyglutamine stretches in the mutant proteins. A separate class of trinucleotide repeat diseases is obtained if the repeat stretch is found in non-coding areas of the gene. This class is not considered further here.

Disease	Host Protein	CAG repeat length
Huntington's disease (HD)	Huntingtin	36-121
Spinocerebellar ataxia type 1 (SCA1)	Ataxin-1	39-83
Spinocerebellar ataxia type 2 (SCA2)	Ataxin-2	32-77
Spinocerebellar ataxia type 3 (SCA3)	Ataxin-3	54-89
Spinocerebellar ataxia type 6 (SCA6)	$Ca_V2.1$-α	19-33
Spinocerebellar ataxia type 7 (SCA7)	Ataxin-7	37-306
Spinocerebellar ataxia type 17 (SCA17)	TATA-BP	47-55
Spinal and bulbar muscular atrophy (SBMA)	Androgen Receptor	40-63
Dentatorubral pallidoluysian atrophy (DRPLA)	Atrophin-1	49-84

Table 1.a: **Overview of the exonic CAG-repeat diseases.**[13] Diseases are listed along with the host protein and the mutant allele repeat number. The wild-type allele repeat numbers are generally non-overlapping ranges of shorter lengths. All these numbers are based on patient data and hence hampered by small sample sizes due to generally low prevalence.[14-16]

All the diseases listed in Table 1.a are hereditary age-of-onset diseases, *i.e.*, symptoms start to develop later in life, usually when patients reach 30-50 years in age. The severity of symptoms is usually progressive, although none of the diseases are directly fatal (see I.2.2).

The host proteins are generally unrelated in both sequence and function (see I.2.3). This finding has dominated the hypotheses formulated with respect to CAG expansion diseases: Pathogenesis is assumed to be triggered by the

unifying characteristic, *viz.* the polyglutamine stretches. Support comes directly from clinical data, which show that for all nine diseases the age-of-onset is inversely correlated to the length of the polyglutamine expansions.[16] The presence of inclusions in the brains of the first mouse model for HD[17] established protein aggregates as a histological hallmark of CAG repeat diseases. This placed them in the broader category of age-of-onset protein aggregation diseases such as Alzheimer's and Parkinson's. When a qualitatively similar, inverse dependence of *in vitro* aggregation rates on repeat length was established with isolated peptides and truncation constructs (see I.2.5),[18,19] it appeared quite reasonable to formulate an overarching, universal hypothesis of the pathogenic mechanism of CAG repeat diseases: The aggregation of protein fragments rich in glutamine and not the details of the host protein and its biology is the crucial pathogenic event.[20,21] This is not dissimilar from the amyloid cascade hypothesis[22] formulated for Alzheimer's disease (also see I.2.6). Additional support for such a hypothesis emerged from several studies *in vivo*, in which somewhat universal behavior was observed even for very disparate sequence constructs (see I.2.4).

In recent years, however, doubt has been cast on the universality of this polyglutamine- and aggregation-centric view. Sequence context is considered with renewed emphasis and the wild-type biology of the host proteins has become a dominant area of research (see I.2.4 and I.2.7).[2,23] Therapeutic strategies focus not just on interfering with protein aggregation but target other implicated cellular pathways as well. It remains to be seen how much merit the

aggregation-centric view will hold for the efficient design of treatment and strategies and ultimately a cure.

I.2.2. Symptoms and Treatment

HD is estimated to be the most prevalent of all CAG-repeat diseases, although reliable numbers have not been established for all the diseases in Table 1.a, and – more importantly – prevalence varies drastically with population. HD was characterized first by the physician George Huntington in the second half of the 19th century by the vivid descriptions of one of its symptoms: chorea, *i.e.*, brief and arrhythmic muscle contractions leading to erratic and uncontrolled motion.[16,24]

Generally speaking, the course of HD and DRPLA can be divided into three stages: i) a pre-symptomatic stage during which patients are completely healthy by clinical standards; ii) a weakly symptomatic stage during which patients might be unaware of any symptoms but careful tests reveal a quantifiable phenotype, and iii) a strongly symptomatic phase which is diagnosed by clinical signatures such as chorea, motor impersistence, or lack of coordination (ataxia). Stage iii) is accompanied by cognitive symptoms; they often include an impairment of cognitive control, *i.e.*, patients have difficulties in organizing, planning, and coordinating tasks. This may lead to an emotional detachment of the individual from her or his social environment and result in an increased risk for suicide.[25] By the neurodegenerative nature of the disease, both motor and cognitive functions are impaired. This makes patient care during advanced stages of the disease a necessity.

SBMA or Kennedy's disease is an X-linked muscle weakness syndrome. In its characterized form, it is found in males only, although female carriers show very mild but quantifiable symptoms.[15,26] Through neurodegeneration, voluntary muscle movements are negatively affected in the limbs, mouth, and throat. This impairs mobility, speech, and the ability to swallow. Symptoms of the disease are relatively mild compared to other CAG repeat diseases and very rarely include cognitive impairment. The primary risk of premature death comes from a weakened respiratory system via secondary infections.[15] Conversely, those spinocerebellar ataxias which are polyglutamine-based (see Table 1.a) exhibit a wide range of symptoms: the most common feature is – as the names of the diseases suggest – a lack of motor coordination resulting in altered gait, posture, and oculomotor deficits.[14] In general, symptoms are highly variable and – unlike SBMA – have considerable overlap with those seen in HD and DRPLA.

Currently, there is no cure for any of the CAG repeat diseases. Because the molecular mechanisms of pathogenesis remain poorly understood, treatment approaches are purely symptomatic. For SBMA, vitamins are administered to help with muscle cramps. The clinically most advanced strategy beyond that has been to reduce the androgen levels in the affected patients.[26,27] Both castration and chemical reduction of testosterone levels improved symptoms and slowed down progression in a mouse model of SBMA.[28] For HD and the ataxias, the situation is similar if not worse. Pharmacological treatment is occasionally reported to be beneficial but remains largely ineffective.[14,16,29] An illustrative data point comes from the observation that the survival expectancy in a remote

Venezulean patient population was very similar to that of populations with ready and thorough access to pharmacological treatment options.[30] Non-pharmacological strategies provide another route to symptomatic treatment. Logopedics, physiotherapy, and counseling are all vital in maintaining as much quality of life as possible for both the patient and her or his family.

I.2.3. Repeat Instability and Host Proteins

The molecular origin of the genetic instability which leads to the expansion of the trinucleotide repeat region on the gene remains poorly understood.[31] In general, we can distinguish between somatic expansion and germline instabilities. The former has been demonstrated for HD[32,33] giving rise – for example – to tissue-specific expansion patterns.[34] This is an important aspect since the disease phenotype is itself tissue-specific suggesting a connection between the two.[35,36] Even more strikingly, it was demonstrated that DNA polymorphism can occur in adult, post-mitotic neurons.[37] Distinct from somatic mutations, germline instabilities create mutations in an intergenerational sense, *i.e.*, meiotic expansion of the trinucleotide repeat during spermatogenesis and oogenesis might predispose future generations to an earlier age-of-onset (a phenomenon referred to as anticipation).

The molecular mechanism for the instability remains somewhat speculative.[38] Roughly speaking, during DNA replication, the trinucleotide repeat on one of the separated strands is predisposed to secondary structure formation, typically a hairpin. This hairpin can occur on the template strand leading to contraction or on the nascent strand leading to expansion. By interfering with

replication, ultimately the trinucleotide repeat stretch self-mutates. If it is located on a coding region, the translated proteins will reflect that mutation. What are the functions, localizations, and sizes of the host proteins such that the organism tolerates such variable mutations?

Table 1.b gives an overview of the disease proteins:

Host Protein	Localization	Function	Size in kD
Huntingtin	Cytoplasma	Associated with diverse cellular pathways / functions	~350
Ataxin-1	Neuronal Nuclei		~90
Ataxin-2	Cytoplasma		variable
Ataxin-3	Cytoplasma	Deubiquitinase[39]	~40
$Ca_V2.1$-α	Cell Membrane	Subunit of a voltage-dependent calcium channel[40]	~282
Ataxin-7	Nucleus	Part of transcriptional regulatory and histone acetylation complexes[41]	~100
TATA-BP	Nucleus	Transcriptional regulator[42]	~35
Androgen Receptor	Nucleus and Cytoplasma	Steroid receptor (testosterone)[43]	~110
Atrophin-1	Cytoplasma	Transcriptional corepressor via nuclear receptors[44]	~125 (83)

Table 1.b: **Overview of host proteins for the nine exonic CAG-repeat diseases.** Proteins are listed (see Table 1.a) along with their typical localization, dominant characterized function, and approximate size. Ataxin-1, Ataxin-3, and Huntingtin have all been vaguely implicated in multiple cellular processes. To an extent, this is true for

Ataxin-3, Ataxin-7, and Atrophin-1 as well. The problem stems largely from the difficult conversion from identified macromolecular interaction partners to assigned function. Sizes were obtained from transcript entries via Ensembl.[45]

As Table 1.b suggests, the native function of the host proteins implies that the most common cellular process implicated in polyglutamine diseases would be transcriptional regulation. In I.2.4, it is argued that a general mechanism not reliant on host protein function emerges. Such a general mechanism is supported by the rather uniform phenotype (see I.2.2) given the unrelated nature of the gene products listed in Table 1.b. No significant sequence similarities and no structural similarities have been discovered for the host proteins.[46]

I.2.4. Suggested Pathogenic Mechanisms

In this section we review the general pathogenic mechanism brought forth to explain the deleterious effects of poylglutamine expansion on neuronal cells. Ironically, this excludes the most obvious explanation, which is as follows: the host protein is adversely affected by the expansion – for example misfolded – and fails to perform its biological function in the same manner as the wild-type protein would. As outlined in I.2.1, this loss-of-function hypothesis by construction postulates a specific mechanism for each disease given that the host proteins are unrelated (also see I.2.3).[23] But not all seemingly specific hypotheses do exclude the possibility of a generic mechanism: An example was touched upon in I.2.2 for SBMA for which the host protein, *viz.* the androgen receptor, allows a specific modulation of the disease phenotype by reducing testosterone levels.[28] This approach is probably indirectly linked to polyglutamine: it appears

reasonable to assume a subsequent down-regulation of the expression levels of the androgen receptor. The reduced amount of disease protein would then consequently reduce the disease phenotype. Another mechanism we do not consider here is that of a direct toxicity of the mutant mRNA construct. Such a role was recently demonstrated for SCA3.[47]

But what cellular pathways do the disease protein, its putative proteolytic fragments, and the aggregation intermediates interfere with? Here, we consider three major processes, which are not necessarily separable:

i. The polyglutamine expansions have been shown to disrupt wild-type protein-protein interaction networks through a coupled loss- and gain-of-function. Native interactions are lost, and new, deleterious interactions are formed by the mutant protein. The polyQ-expanded host protein of SCA1, ataxin-1, has been demonstrated to interact more favorably with a putative RNA-binding protein but less favorably with a transcriptional repressor protein.[48,49] In this case, the affected downstream process is almost certainly transcriptional regulation. It is altered through an upstream modification of protein-protein interactions which are directly polyglutamine-dependent. Similarly, the host protein for SCA17, *viz.* the transcriptional regulator TATA-box binding protein (TBP), has been shown to exhibit a reduced propensity to homodimerize in a polyglutamine length-dependent fashion. Similar to ataxin-1, an interaction with a different transcriptional regulator is positively affected leading to a direct hypothesis for pathogenesis through altered transcriptional regulation.[50] Interestingly, TBP has also been shown to be sequestered in Huntingtin-

containing aggregates through a presumed interaction of the polyglutamine stretches.[51] Lastly, polyQ-expanded TBP also shows reduced interactions with its other native binding partner, DNA.[52]

ii. The formation of larger oligomeric but soluble species might operate as a generic modulator of cellular function by engaging in degenerate protein-protein interactions. Vital proteins might be sequestered into growing oligomers causing stress to the cell.[53] This is a direct generalization of i), but focuses on non-specific interactions mediated by the polyglutamine tract. Indirect support for such a mechanism comes from studies *in vivo* employing overexpressed constructs with polyQ-expansions beyond the pathological length threshold. Takahashi *et al.*[54] demonstrated that sequence constructs with polyglutamine stretches attached to truncated native sequences derived from Huntingtin or atrophin-1 and additionally tagged with GFP form soluble oligomers, which are cytotoxic in a length-dependent manner. Similarly, Wong *et al.*[55] characterized SDS-insoluble, spherical oligomers in an inducible *Drosophila* model of SCA3. The presence of these oligomers was shown to correlate with the neurodegenerative symptoms exhibited by the fly model. Such a generic model of toxicity is also supported by the existence of a common antibody recognizing amyloid-like oligomers.[56] Both of the aforementioned studies[54,55] provide further evidence that microscopic, precipitated aggregates in cells (inclusion bodies) are not correlated with neurodegeneration and cell survival. A controversial point several years ago, enough evidence has been brought forth that it now seems widely agreed

upon that visible, cellular aggregates are circumstantial signs of neurodegeneration but are in no way causative.[23,57-59]

iii. As a direct consequence of both i) and ii), polyglutamine expansions might impair the cellular protein quality control system (PQCS). Heat shock proteins in the 70kD family (Hsp70) have been detected in or associated with cellular aggregates induced by pathological polyQ-expansions.[53,60,61] Proteins in the Hsp70 family often act as molecular chaperones and are hence partially responsible for the folding and re-folding of nascent and misfolded proteins. The molecular chaperone machinery is linked through regulatory interaction networks to the ubiquitin-proteasome degradation pathway. This linkage includes a few well-characterized proteins such as CHIP or BAG1. Upregulation of proteins in or associated with the Hsp70-family can be protective[62-65] although the complexity of the network might prevent the overexpression of just a single protein from having any effect. A universal mechanism of pathogenesis is suggested, however, by the multiple independent findings that components in the PQCS reduce polyglutamine-induced toxicity upon overexpression or exacerbate it upon siRNA suppression. Recent cases include but are not limited to such demonstrations for the regulatory protein sacsin in a model of SCA1,[66] CHIP in a model of SCA3,[67] and p97-type chaperones,[68] the TRiC chaperonin,[69,70] and HYPK[71] in models of HD. A model for pathogenesis emerges in which the mutant protein incapacitates the PQCS by resisting both proteolytic degradation and

chaperone-assisted refolding, which ultimately has fatal downstream consequences for the cell.[72-74]

Other pathways and systems which have – at least temporarily – attracted a considerable amount of attention involve mitochondrial function[75,76] and protein, in particular histone acetylation.[77]

However, it is not at all straightforward to identify a causative role for aggregation in the above mechanisms. While the examples listed in ii) demonstrate correlation between neurotoxicity and the presence of soluble aggregates, more direct evidence for a causative effects comes from the protective effects exhibited by aggregation inhibitors: early work including small molecules such as trehalose[78] was performed exclusively *in vivo* and hence remains indirect. Since then, compounds such as polyphenols found in green tea[79], proline[80], and the amlyoid-specific dye Congo Red[81,82] have all been shown to inhibit aggregation *in vitro*, and to alter pathogenesis in disease models. Most recently, renewed interest has been shown in the undecapeptide QBP1,[83] which had been identified and characterized as a peptidic aggregation inhibitor.[84] All these studies suggest that the process of protein aggregation remains at least partially responsible for pathogenesis in exonic CAG repeat diseases.

One area that remains the subject of much scrutiny is the role of proteolytic cleavage.[85-89] Analysis is hampered by the fact that it is very difficult to identify prominent fragments *in vivo*. Several studies suggest that suppression of proteolysis reduces toxicity.[90,91] This gives rise to the “toxic fragment” hypothesis, which is one of the major justifications of studying the intrinsic properties of the

polyglutamine stretch alone. Figure 1.1 summarizes the above discussion as a graphical sketch:

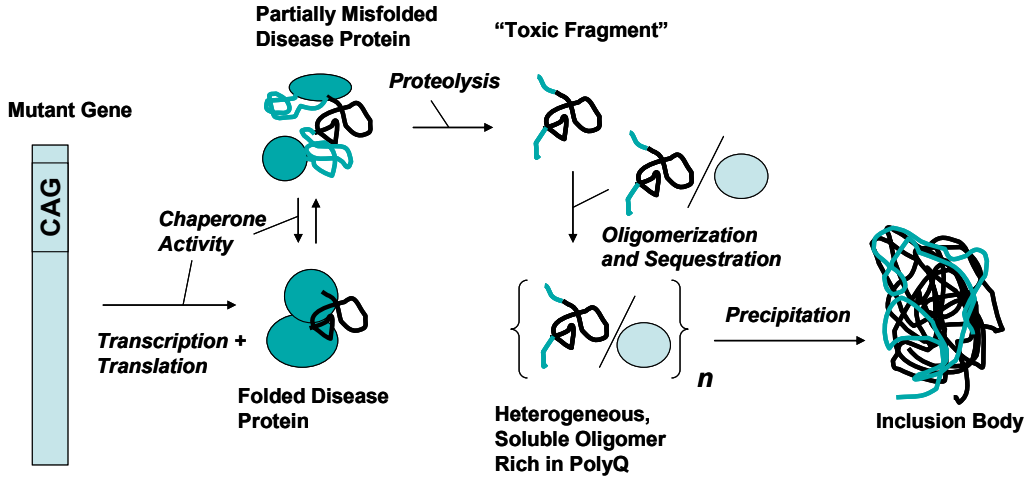

Figure 1.1: **A graphical illustration of the possible progression of CAG repeat disease pathology**. The expanded gene yields a mutant protein. The polyQ-expansion is generally unstructured and might cause the host protein to partially unfold. The PQCS is constantly occupied with remedial activity. Eventually, a proteolytic fragment is obtained, which oligomerizes heterogeneously. A size-threshold is passed and a cellular precipitate (microaggregate or inclusion body) forms, which might have amyloid-like characteristics. The PQCS remains stressed by continuing efforts to clear away "misbehaved" protein material. Cell death occurs, although it is not known which process ultimately triggers it, nor whether there even is a universally applicable mechanism.

Aggregation – a key event in Figure 1.1 – has been characterized extensively for model peptides *in vitro*, and this is discussed next.

I.2.5. Aggregation Studies and Kinetic Analysis

Peptides rich in glutamine are prone to aggregation (see I.2.6). This inherent tendency has complicated the *in vitro* analysis of this system due to the

difficulty in preparing an aggregate-free sample. Early studies by Scherzinger *et al*.[18] revealed a striking dependence on the polyglutamine stretch length (N) for the aggregation of a GST-based protein construct containing flanking sequences of the Huntingtin protein (see Table 1.c). The rate of aggregation increased strongly with N and with concentration. Moreover, the authors showed the aggregation data to be consistent with a nucleation-dependent mechanism. The most obvious demonstration came from the fact that adding pre-formed aggregates completely eliminated the kinetic lag-phase.

Wetzel and colleagues then established protocols to reliably perform *in vitro* aggregation assays using more reduced model systems, *i.e.*, synthetic peptides composed entirely of glutamine except for two lysine residues on either end for increased solubility ($K_2Q_NK_2$). Protocols to overcome technical problems in the purification, storage, and disaggregation of these synthetic peptides were developed.[92] Based on the work of Ferrone,[93] *in vitro* aggregation data were analyzed as follows. Consider a generic, step-wise polymerization reaction for species A:[94]

$$\begin{aligned} A + A &\rightleftharpoons A_2 \qquad K_2 = \frac{k_2}{k_{-2}} = \frac{[A_2]}{[A]^2} \\ A_i + A &\rightleftharpoons A_{i+1} \qquad K_{i+1} = \frac{k_{i+1}}{k_{-(i+1)}} = \frac{[A_{i+1}]}{[A_i][A]} \end{aligned} \tag{1-1}$$

In Equation 1-1, K indicates equilibrium constants, k rate constants, and square brackets denote activities. Only the initial (dimerization) step and the

generalization for later steps are shown. For a nucleated process, let us assume a rapid pre-equilibrium controlling the association up to a nucleus of size n^* :

$$n^* A \rightleftharpoons A_{n^*} \qquad K_{n^*} = \frac{k_{n^*}}{k_{-n^*}} = \frac{[A_{n^*}]}{[A]^{n^*}} \tag{1-2}$$

We can now define the net rate of formation of growing polymer ends by:

$$\frac{dc_p}{dt} = k_+^* c_* [A] - k_-^* c_* = K_{n^*} \cdot [A]^{n^*} \cdot \left(k_+^* [A] - k_-^* \right) \tag{1-3}$$

In Equation 1-3, c_* denotes the concentration of nuclei and c_p the concentration of growing ends. The pre-equilibrium in Equation 1-2 was used to obtain an expression in powers of $[A]$. If we assume aggregate size-independent rate constants beyond the nucleus size n^*, then monomer loss is governed by:

$$\frac{d\left(c_t - [A]\right)}{dt} = \frac{d\Delta}{dt} = c_p \cdot \left(k_+ [A] - k_- \right) \tag{1-4}$$

Here, c_t is the total monomer concentration. We make the following further assumptions:

$$k_-^* = k_- \approx 0 \quad ; \quad k_+^* = k_+ \quad ; \quad [A] \approx c_t = const. \tag{1-5}$$

The assumption about the free monomer concentration essentially corresponds to a focus on the *initial* rate of polymerization. We can then combine Equations 1-3 and 1-4 to yield:

$$\frac{d^2\Delta}{dt^2} = \frac{d}{dt}\left(c_p k_+ c_t \right) = k_+ c_t \frac{dc_p}{dt} = K^{n^*} k_+^{\,2} c_t^{\,n^*+2} \tag{1-6}$$

Equation 1-6 can be integrated to give a very rough estimate of the initial time course of a polymerization reaction in which a well-defined nucleation event is followed by irreversible, kinetically uniform, downhill addition of monomers:[19,93]

$$\Delta(t) = \frac{1}{2} K^{n*} k_+^{\,2} c_t^{\,n*+2} t^2 \tag{1-7}$$

By plotting Δ as a function of t^2, the pre-factor containing the two relevant constants can be obtained via linear regression.[19] Measurement of this pre-factor as a function of concentration allows the determination of the nucleus size, $n*$, by using a double logarithmic plot:

$$\frac{d \ln\left(\frac{d^2\Delta}{dt^2}\right)}{d \ln c_t} = \left(n*+2\right) \cdot K^{n*} k_+^{\,2} c_t^{\,n*+1} \cdot \frac{c_t}{K^{n*} k_+^{\,2} c_t^{\,n*+2}} = n*+2 \tag{1-8}$$

Furthermore, Wetzel and co-workers developed an assay which allowed the quantification of c_p when the nucleation rate is slow in comparison.[95] Then, monomer loss can be measured to yield an effective first-order rate constant according to Equation 1-4 with the assumption that c_p is constant. From this, the elongation rate constant k_+ may be extracted which in turn allows the determination of K^{n*} from a slope of a plot of Δ as a function of t^2.

In 1962, Oosawa and Kasai[96] proposed an alternative simplification for homogeneous nucleation in which nucleus formation is treated as an irreversible, kinetic event. The resultant expression is:

$$\ln\frac{1+x}{1-x} = 2\cdot\sqrt{k^{n*}k_{+}nc_{t}^{\ n*}}\cdot t$$
$$x = \sqrt{1-\left(\frac{[A]}{c_t}\right)^{n*}} \qquad (1\text{-}9)$$

In Equation 1-9, k^{n*} is the effective rate of formation of nuclei. A series expansion of the logarithmic term yields that the leading dependence of Δ is on t^2 which is in agreement with the approximation shown in Equation 1-7. The models can be matched up by the following proportionality:

$$k^{n*} \propto c_t k_{+} K^{n*} \qquad (1\text{-}10)$$

Using Equations 1-9 and 1-10, we obtain simulated data over the entire time course and are able to test the robustness of the analysis outlined in Equations 1-7 and 1-8 given that an accurate determination of the actual, *initial* rate of aggregation may be very difficult if not infeasible. Of course, the Oosawa-Kasai model still maintains the fundamental tenet that the process is homogeneously nucleated. If we find that the resultant estimates for n* depend astutely on the approximations introduced, then experimental errors may be large and a re-analysis using a more complete model of aggregation is in order. If we, however, find that the estimate of n* is fairly robust, then we can infer something about the underlying aggregation mechanism should – for example – estimates of n* be fractional rather than integer numbers (see below and Chapter VII).

Figure 1.2 shows simulated data for the initial time course using the models as given by Equations 1-7 and 1-9 for different concentrations c_t. We use values reported in the literature of $k_+ = 10^4 M^{-1} s^{-1}$, $K^{n^*} = 2.6 \cdot 10^{-9}$, and $n^* = 1$.[95]

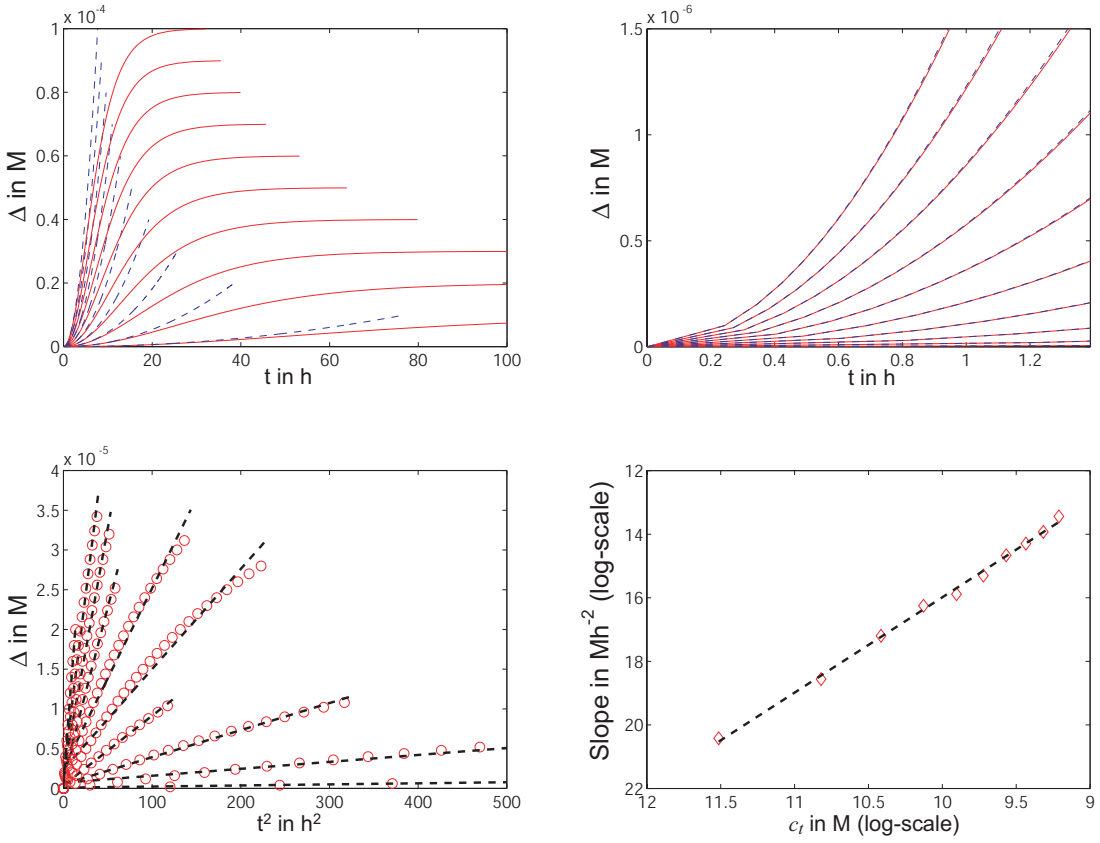

Figure 1.2: **Simulated kinetic aggregation data**. Panel A shows the entire course of aggregation according to Equation 1-7 (blue dashed lines) and Equation 1-9 (red lines). Panel B shows a close-up of the initial stage of aggregation. Clearly, the leading dependency on t^2 is shown by both models. Panel C plots the aggregation time courses from Equation 1-9 against t^2 (red circles). Deviations from linearity are minor as indicated by linear fits (black, dashed lines). Panel D is a double-logarithmic plot of the slopes of the lines in Panel C against total concentration. Panels C and D are equivalent to the analysis by Chen *et al*. and the slope in Panel D yields the critical nucleus size.

As Figure 1.2 points out (see below), the t^2-model is very limited in describing the long-time behavior of aggregation in accordance with expectation (Panel A) but agrees well with the Oosawa model for the initial phase (Panel B). In order to evaluate the robustness of the analysis yielding the critical nucleus size, we introduce substantial noise. In Panel C, linear fits according to Equation 1-7 are shown (Δ vs. t^2) with the data obtained from the (more realistic) Oosawa model. Since in experimental work the ranges used to define the "initial" phase are not rigorously controlled, we sample the cutoff uniformly in an interval from 10-60% of monomer loss. This is the source of considerable noise as indicated by the varying quality of the linear fits in Panel C. Even then, the nucleus size is extremely well-described by the linearization shown in Panel D. The slope of the line is ~3.0 in perfect agreement with the expectation value of 3.0 (Equation 1-8). Repetition showed that the noise in the estimate is small and normally distributed with a standard deviation of about 0.06 (data not shown). More importantly, if the arbitrariness of defining the initial phase is removed, the fit in Panel D becomes independent of the quality of the linear approximation in Panel C, and the slope is always *exactly* 3.0 (data not shown).

Wetzel and co-workers used the model defined in Equation 1-7 and partially tested in Figure 1.2 to conclude the following about the aggregation of polyglutamine-based synthetic polypeptides *in vitro*:

- The aggregation follows the mechanism of homogeneous nucleation. This is weakly established by the linearity of plots of Δ vs. t^2. Such a dependence is by no means a unique characteristic of a nucleated polymerization

reaction.[93,97] More – but equally suspect[97] – evidence comes from the qualitative observation of the ability to seed the aggregation, *i.e.*, to attenuate the lag-phase by the addition of a small amount of pre-formed aggregates.[19,98]

- The size of the nucleus obtained via Equation 1-8 is less than unity (equivalent to the slope in Panel D of Figure 1.2 being significantly less than three).[19] If homogeneous nucleation applies, this result appears only to be consistent with a monomeric nucleus. In this scenario, Equation 1-2 relaxes to the equilibrium for a critical conformational event. The peptide undergoes a rare transition[95] to yield a toxic species which is prone to aggregation.
- The aggregation kinetics are chain length-dependent. The longer the peptide the faster aggregation occurs. The exact dependencies of the constant terms in Equation 1-7 are not known. For $K_2Q_{47}K_2$ elongation was estimated to be significantly slower than a diffusion-limited reaction would suggest.[95]
- The pre-equilibrium constant K^{n^*} is very small and corresponds to a free energy barrier of more than 12 kcal/mol for the peptide $K_2Q_{47}K_2$. This suggests that free nuclei would never be observed in solution at measurable concentrations. The barrier is expected to decrease with increasing chain length and vice versa.[95]

Figure 1.3 sketches the proposed mechanism graphically. The monomeric nucleation event is interpreted and depicted as a conformational transition, specifically a disorder-to-order transition. However, several lines of criticism may be formulated against this simple picture.

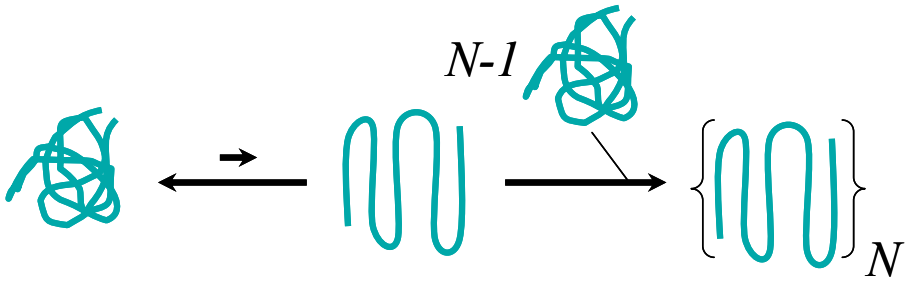

Figure 1.3: **The homogeneous nucleation model for polyglutamine aggregation**. Disordered, soluble polyglutamine undergoes a rare folding transition to a "toxic" conformer with equilibrium constant K^{n^*}. This monomeric nucleus seeds the downhill aggregation to the mature, fibrillar aggregate via monomer addition. The elongation rate is too slow to be consistent with a diffusion-limited reaction, and hence a "dock-and-lock" mechanism is proposed: initial docking is followed by slow conformational re-arrangement. Most of the aggregation rate dependence on chain length is hypothesized to be captured by the first step.

- The nature of the conformational transition is unknown. The assignment of a folding transition from a disordered species to a β-rich conformer is based entirely on the known facts about the endpoints of the reaction (see I.2.6). The fully mature fibril, however, is a species so far removed from the monomer that this assignment seems questionable. Evidence is usually seen in the fact that different experimental metrics of monomer loss (such has HPLC) coincide on a coarse time schedule with the formation of β-secondary structure as determined by CD spectroscopy.[19]
- Rigorously speaking, the kinetic model implies the absence of soluble intermediates, and this has been argued to be the case.[19] There is considerable evidence, however, that soluble oligomers form and are valid

reaction intermediates.[54,99-102] This would much more likely be consistent with a heterogeneous reaction in which multiple pathways for aggregation and/or multiple free energy barriers along the aggregation pathway(s) exist.[103] Such a model appears particularly relevant in an *in vivo* setting.

- The linear fits which yield the nucleus size were shown to be very robust in theory (see Figure 1.2). Given that the obtained values for n^* are typically less than unity,[19] it is reasonable to conclude that the model does not apply rigorously. In accordance with the previous point, such significant deviations from unity suggest the need for a more complicated model even *in vitro*. This finding is exacerbated in the presence of other flanking sequences.[104]

It is very important to emphasize that the use of a model which fits the data reasonably well is in itself not a matter of any concern. Rather, the mechanistic interpretation of such a model based simply on that fact that it *does* apply merits further scrutiny.[97] It is one of the core aims of this thesis to evaluate the plausibility of the above model using tools which are not inherently limited by their inability to detect and characterize small oligomers,[105] *viz*. computer simulation (see I.3 and Chapters IV, V, and VI).

In silico work on the issue of protein aggregation has been invaluable in providing a molecular picture associated with the process. Dima and Thirumalai[106] showed that protein folding and aggregation are inherently coupled and that the resultant phase diagram is much richer than the simple kinetic analysis outlined above (which implies only two distinct phases) would suggest. Using simplified models, nucleation-dependent aggregation mechanisms have

been established as feasible models and their dependency on environmental and intrinsic parameters has been investigated.[107-109] As might be expected, such studies are invaluable in establishing a quantitative, conceptual framework for defining the problem but sacrifice realism for feasibility.

I.2.6. Structural Characteristics of Polyglutamine

Monomeric, Soluble Polyglutamine

One of the dominant characteristics of soluble polyglutamine is the absence of a consensus structural signature. Both synthetic polypeptides and even polyQ-expansions in artificial or native host systems are universally characterized as being disordered. This result is obtained predominantly by the use of CD spectroscopy[19,98,110-112]. Further corroboration comes from NMR experiments[111] as well as computer simulation[113] including this thesis.[5,7]

The lack of a consensus structure has led to the unfortunate notion that polyglutamine is random coil-like in solution. This has certainly been interpreted to imply that the peptides populate ensembles of swollen conformations without any further proof. It was subsequently established via FCS[114] that this notion is in fact incorrect. In water, polylgutamine-based peptides populate collapsed structures whose hydrodynamic radii grow with chain length in a way consistent with water being a poor solvent for this system (see Chapter II). In a poor solvent, chain-chain interactions are preferred over chain-solvent interactions, and the dominant conformations are disordered, collapsed globules. This result is in congruence with the well-known insolubility of synthetic polymers[115] rich in

glutamine and suggests a generic driving force for both collapse and aggregation.[103]

It should be pointed out that there is a minimum length scale for this behavior, which is given by the inherent stiffness of polypeptides. The concept of the "blob length", *i.e.*, the length scale over which the inherent polymeric behavior is masked by conformational rigidity,[116] may be used to explain why a recent experimental study found triplet contact quenching data (Trp-Cys) to be consistent with extended conformations for short, synthetic peptides composed predominantly of glutamine.[117] Similarly, a crystal structure has been published showing a short, glutamine-based peptide bound to a polyQ-antibody.[118] At present, it is difficult to adjudicate how relevant this structure is for the conformational ensemble of polyglutamine in solution due to similar concerns.

Polyglutamine Aggregates

The finding that aggregates derived from mutant Huntingtin do in fact show amyloid-like features[119,120] was one of the key results placing HD and other CAG repeat diseases in the context of prominent amlyoidoses like AD. Unfortunately, the term amyloid is phenomenologically defined and relies on specific experimental signatures of protein aggregates[121] such as: i) Congo red staining and birefringence under polarized light; ii) CD and fiber diffraction signals indicative of β-secondary structure; iii) thioflavin T (ThT) binding and fluorescence shift; iv) fibrillar architecture with characteristic dimensions and twist as seen in electron micrographs. While polyQ-derived fibrils exhibit some of

those features, they typically do not or only weakly exhibit the characteristic Congo red birefringence.[120] The most common, amyloid-specific probe of fibril formation for polyglutamine-based peptides in *in vitro* experiments is ThT fluorescence. Other methods that do not rely on structural characteristics include light scattering and quantification of the soluble fraction via HPLC. Coincidence of these methods has been used to argue against the presence of disordered, insoluble intermediates and aggregates.[19]

Previously, Perutz had speculated that polyglutamine would be amenable to an amyloid-like arrangement,[122] which is characterized by the presence of ordered β-strands perpendicular to the fiber axis. Two structural models were derived from the diffraction data: the prominent nanotube-like β-helix[123] and the multi-pleated sheet[124,125]. Much computational effort has been invested to create and test structural models of assembly of polyQ-based peptides in a fashion consistent with either of these two possibilities. Such studies are ultimately limited by their inability to probe the energy landscape beyond the local minimum the system was initially prepared in; hence, results have been obtained which are either conflicting or point toward the possibility of substantial heterogeneity in assembly structures.[126-131]

I.2.7. Sequence Context Dependencies

As was touched upon in I.2.4, the issue of proteolytic cleavage remains an active point of investigation. It is presumed that fragments rich in glutamine are dominant toxic players, but how are their properties modulated by the wild-type

sequence context? *In vivo*, it is nearly inevitable that some amount of variation occurs with the ongoing processes of proteolytic degradation and heterogeneous association (see Figure 1.1). Considerable evidence has been brought forth in *in vitro* experiments that sequence context does matter but that it can be masked by the inherent properties of the polyQ-expansion. Robertson *et al*.[112] showed that terminal polyQ-expansions did not alter the properties of a model host protein and that the host protein did not alter the qualitative aggregation behavior mediated by the polyQ-expansion. Conversely, if the polyglutamine stretch was moved to the interior of the protein, the host protein was drastically destabilized. Nagai *et al.*[132] found that they could isolate a monomeric form of a fusion protein of thioredoxin and polyglutamine, which was shown to act qualitatively as an aggregation seed. The authors used CD to show that the toxic form is β-rich via equilibrium measurements. This result clearly is a function of the artificial sequence context and could not have been obtained with peptide constructs such as those used by Bhattacharyya *et al*.[95]

It might therefore seem questionable to employ *de novo* sequence constructs to study the biomedically relevant properties of polyglutamine. Table 1.c gives an overview of the wild-type sequence context for the nine exonic CAG repeat diseases:

Host Protein	**Gene (MIM)**	**Sequence**
Huntingtin	143100	MATLEKLMKAFESLKSF**Q**$_N$PPPPPPPPPPPQLPQ PPPQAQPLLPQPQPPPPPPPPPPGPAVAEPLHRP KKELSATKKDRVNHCLTICENIVAQSVRNS

Host Protein	Gene (MIM)	Sequence
Ataxin-1	164400	AHLPHTFQFIGSSQYSGTYASFIPSQLIPPTANPVTSAVA SAAGATTPSQRSQLEAYSTLLANMGSLSQTPGHKAE**Q**$_N$HL SRAPGLITPGSPPPAQQNQYVHISSSPQNTGRTASPPAIPV HLHPHQTMIPHTLTLGPPSQVVMQYADSGSHF
Ataxin-2	183090	PTRASPLGARASPPRSGVSLARPAPGCPRPACEPVYGPLT MSLKP**Q**$_N$PPPAAANVRKPGGSGLLASPAAAPSPSSSSVS SSSATAPSSVVAATSGGGRP
Ataxin-3	109150	DMEDEEADLRRAIQLSMQGSSRNISQDMTQTSGTNLTSEE LRKRREAYFEKQQQK**Q**$_N$GDLSGQSSHPCERPATSSGALGS DLGDAMSEEDMLQAAVTMSLETVRNDLKTEGKK
$Ca_V2.1$-α	183086	GTSTPRRGRRQLPQTPSTPRPHVSYSPVIRKAGGSGPP**Q**$_N$A VARPGRAATSGPRRYPGPTAEPLAGDRPPTGGHSSGRS PRMERRVPGPARSESPRACRHGGARWPASG
Ataxin-7	164500	MSERAADDVRGEPRRAAAAAGGAAAAAAR**Q**$_N$PPPPQPQRQ QHPPPPPRRTRPEDGGPGAASTSAAAMATVGERRPLPSPEV MLGQSWNLWVEASKLPGKDGTELDESFKEFG
TATA-BP	607136	MDQNNSLPPYAQGLASPQGAMTPGIPIFSPMMPYGTGLTP QPIQNTNSLSILEEQQR**Q**$_N$AVAAAAVQQSTSQQATQGTSG QAPQ
Androgen Receptor	313200	MEVQLGLGRVYPRPPSKTYR GAFQNLFQSVREVIQNPGPR HPEAASAAPPGASLLLL**Q**$_N$ETSPRQQQQQQGEDGSPQAH RRGPTGYLVLDEEQQPSQPQ
Atrophin-1	125370	PASSSAPAPPMRFPYSSSSSSSAAASSSSSSSSSSASPFP ASQALPSYPHSFPPPTSLSVSNQPPKYTQPSLPSQAVWSQ GPPPPPPYGRLLANSNAHPGPFPPSTGAQSTAHPPVSTHHH HH**Q**$_N$HHGNSGPPPPGAFPHPLEGGSSHHAHPYAMSPSLGSLR

Table 1.c: **Protein sequences surrounding polyglutamine stretch in host proteins of CAG repeat diseases**. Protein names are listed along with gene access codes for the MIM database. The third column lists the protein sequences for the host protein in the vicinity of the polyglutamine stretch indicates as **Q**$_N$. Data were originally compiled by Tim E. Williamson.

Table 1.c shows that the most dominant unifying feature is the presence of other low complexity sequences in the immediate vicinity of the polyQ-expansion.

Polyserine, polyproline or polyalanine stretches are found in several of the proteins. This suggests that the polyglutamine stretch might inherently be located in a disordered region of the protein, which constitutes a straightforward explanation as to why the cell seems to tolerate such genetic instabilities at all (see I.2.3). This finding correlates somewhat with the pattern observed for the *maximum* mutant repeat lengths shown in Table 1.a, which are lowest for the gene products with well-known function (TATA-BP, androgen receptor, and $Ca_V2.1$-α, see Table 1.b).

Most careful proteolytic analyses have focused on huntingtin (htt).[133,134] The sequence context of this protein has also triggered the most studies employing artificial constructs mimicking what a wild-type fragment could putatively look like. In htt, the polyQ-expansion is close to the start codon for exon1. The effect of the 17 residues N-terminal of the polyQ-expansion has been recently investigated. *In vitro* aggregation experiments suggest a substantial enhancement of the aggregation propensity for $htt^{N\text{-}T}Q_N$ with respect to the $K_2Q_NK_2$ constructs used before. It was hypothesized that the hydrophobic residues in the N-terminal fragment enhance the aggregation propensity and rate.[104] Experiments *in vivo* showed that the presence of the 17 N-terminal residues alters the subcellular localization of a fusion construct, usually with GFP.[135,136] On the C-terminal side, the polyQ-expansion is flanked by a proline-rich region. Two *in vitro*-studies have quantified the effect of C-terminal oligoproline flanking sequences on the aggregation of polyglutamine peptides.[137,138] In both cases, it was found that aggregation is retarded and that

the proline segment somewhat disrupts the characteristic structural signatures associated with the growing aggregate (see I.2.6). The effects of flanking sequences on the molecular properties of polyglutamine are the focus of Chapter VI.

I.3. Synopsis

As was suggested in I.2.4, the physicochemical process of aggregation of polyglutamine appears to be a crucial determinant of pathogenesis in exonic CAG repeat diseases. Even though the biological and chemical contexts are known to modulate its behavior, the polyglutamine expansion seems to possess *inherent* properties which mediate and control the process of aggregation. However, detailed, structural models for the early stages of aggregation - including nucleation - are missing due to the inability of present experimental technologies to monitor these low likelihood species.[105] *In silico* work has attempted to remedy this shortcoming. Computer simulations control the resolution directly through the chosen representation of the system and low-likelihood events can be unmasked within equilibrium measurements.

In this thesis, we address several questions regarding structural and mechanistic aspects of polyglutamine aggregation. In detail, the major points are as follows:

i. In Chapter II,[5] we show that concepts borrowed from polymer physics can be used to rigorously quantify the monomeric, disordered state of polyglutamine peptides in solution. We confirm the experimental result that water is a poor

solvent for polyglutamine. These peptides adopt collapsed, globular conformations with no significant consensus structure. This type of disorder has important consequences for the intrinsic dynamics of these peptides. Conformational rearrangement on the collapsed manifold is slow, and glassy behavior is obtained on the nanosecond timescale. The glassiness renders canonical MD approaches employing an explicit representation of the solvent infeasible. Water's poor solvent nature might appear surprising given that the molecules are composed exclusively of polar moieties. However, it is consistent with polylgutamine's aggregation propensity observed *in vivo* and *in vitro* and identifies a generic driving force linking conformational ensembles at the monomer level to the observed driving force for phase separation.

ii. Motivated by the infeasibility of the sampling approach employed in Chapter II to model the phenomena of interest to us, we developed a novel continuum solvation model termed ABSINTH (for self-**A**ssembly of **B**iomolecules **S**tudied by an **I**mplicit, **N**ovel, and **T**unable **H**amiltonian).[6] Chapter III reports on the theory underlying the model and its calibration. It was specifically designed to provide a simple and accurate framework for describing solvation of biomolecules sampled via the Metropolis Monte Carlo method. The degrees of freedom are not the Cartesian coordinates of all atoms but the "essential", internal degrees of freedom, *i.e.*, the dihedral angles along freely rotatable and semi-rigid bonds. We show that ABSINTH accomplishes the difficult task of balancing the strength of specific interactions and structural propensities as a function of simulation temperature. We do so by investigating the model's

ability to reproduce experimentally determined melting profiles for a variety of small polypeptides capable of adopting a specific fold in aqueous solution. Its favorable performance in this regard indicates that ABSINTH is well-suited to describe intrinsically disordered systems such as polyglutamine, which might only transiently adopt specific structural motifs. We provide evidence that there are no qualitative differences between the conformational ensembles for polyglutamine generated by the (presumably) more accurate approach in Chapter II and the novel, implicit solvation approach and that all the results for polyglutamine are in fact consistent with experimental data obtained for soluble polyglutamine.

iii. In Chapter IV,[7] we apply the newly developed continuum solvation model introduced in Chapter III to study the chain length- and solvent quality-dependent properties of polyglutamine at the monomer and dimer levels. We employ simulation temperature as a generic dial for solvent quality: the higher the temperature, the better the solvent. Analysis of globule-to-coil transitions allows us to identify the θ-temperature for this model, *i.e.*, the temperature at which solvent quality is such that chain-chain and chain-solvent interactions are effectively balanced. We can therefore establish a phase diagram for the system by tuning conditions to range from poor via indifferent to good solvent qualities. We find that these molecules spontaneously associate to form stable dimers in the poor solvent regime. This happens at concentrations and environmental conditions approaching those of typical *in vitro* experiments for chains of length $N \geq 15$. The spontaneity of these homotypic associations

increases with increasing chain length and decreases with increasing temperature. Similar and generic driving forces govern both collapse and spontaneous homodimerization of polyglutamine in aqueous milieus. Collapse and dimerization maximize self-interactions and reduce the interface between polyglutamine molecules and the surrounding solvent. There do not appear to be any specific structural requirements for either chain collapse or chain dimerization, *i.e.*, both collapse and dimerization are non-specific in that disordered globules form disordered dimers. These results suggest that polyglutamine aggregation is unlikely to follow a homogeneous nucleation mechanism with the monomer as the critical nucleus. Although we do not test this directly, our results support the formation of disordered and soluble oligomers as early intermediates – a proposal that is congruent with a growing body of experimental data (see I.2.4 and I.2.5).

iv. To confirm the results obtained in Chapter IV, we attempt to directly test the homogeneous nucleation model brought forth for the aggregation of polyglutamine (see I.2.5 and Figure 1.3 in particular) in Chapter V.[8] In particular, we test three of its major tenets:

   a. Is the formation of a putative, monomeric nucleus rich in β-secondary structure an extremely rare event? We quantify the thermodynamic likelihood of polyglutamine to adopt structures consistent with a high amount of β-secondary structure. We employ a biased sampling approach to construct the free energy profile along a reaction coordinate quantifying the

net amount of β-content and find that a putative nucleus would indeed be an extremely low likelihood species.

b. Does the likelihood to form the monomeric nucleus increase with increasing chain length? By performing calculations as detailed in a) for a range of chain lengths spanning the pathological threshold region, we show that the adoption of structures high in β-content becomes *less* likely with increasing chain length. Moreover, we identify no local minima along the reaction coordinate indicating that the putative nucleus would strictly be a transient species.

c. Is the dimerization propensity of molecules pre-organized to resemble putative nuclei significantly enhanced in comparison to disordered conformers? By studying the dimerization propensity in the presence of a structural bias predisposing polyglutamine to adopt β-rich structures, we establish that the presence of canonical β-secondary structure does *not* enhance associativity. We show that, even though the conformational ensemble of individual molecules is drastically altered, the chains associate spontaneously just as much as disordered, collapsed globules do at the dimer level. We do find that β-secondary structure is enhanced at the dimer interface. This suggests that high β-content is a property associated with larger aggregates, for which the collapse constraints imposed by the poor solvent nature of the environment are relieved by complete sequestration from the solvent.

v. Chapter V allowed us to arrive at the conclusion that β-secondary structure plays a role only during the later stages of polyglutamine aggregation. But how are the early stages, *i.e.*, the formation of disordered oligomers, modulated in the presence of flanking sequences? Chapter VI presents work similar in approach to the work presented in Chapter IV: here, we quantify the conformational ensembles and the associativity exhibited by sequence constructs of the 17 N-terminal residues of exon1 of htt (see I.2.7) and polyglutamine stretches of varying length. We show results that establish a significant propensity of the N-terminal fragment to adopt α-helical conformations. Longer polyglutamine stretches suppress this propensity and induce more extended conformations in the N-terminal segment. Overall, the peptides remain disordered and their dimerization is mediated by the polyQ-expansion. Analysis of contact patterns and visual inspection of central structures of dominant clusters reveal that the interactions of hydrophobic residues in the N-terminal fragment are frustrated due to their proximity to charged residues and contribute little to the intermolecular interface. We show that the interfacial penalty experienced by the chimeric peptides is nearly obliterated and that associativity is consequently reduced relative to the homopolymer. However, it appears to be *increased* relative to sequence constructs of the type $K_2Q_NK_2$. Overall, we find a profound alteration of the intrinsic properties of polyglutamine due to the presence of the flanking sequence and comment on consequences for possible pathogenic mechanisms in HD.

The above synopsis focuses predominantly on the context of CAG repeat diseases. Accomplishments outside of this scope are discussed along with the disease-relevant results in Chapter VII.

I.4. Bibliography

A brief note is in order: the bibliography to this chapter is rather exhaustive and covers references relevant to the entirety of this thesis. However, given the variable scope of each chapter, we have decided to include bibliographies specific to each chapter as well. While overlap between cited works from chapter to chapter is inevitable, we made this choice in the interests of clarity and coherence. References are presented in a very condensed format so as to not unnecessarily bloat the written document.

1. Cummings, C. J.; Zoghbi, H. Y. Hum Mol Genet 2000, 9(6), 909-916.
2. Williams, A. J.; Paulson, H. L. Trends Neurosci 2008, 31(10), 521-528.
3. Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. Annu Rev Phys Chem 1997, 48(1), 545-600.
4. Thirumalai, D.; Klimov, D. K. Curr Opin Struct Biol 1999, 9(2), 197-207.
5. Vitalis, A.; Wang, X.; Pappu, R. V. Biophys J 2007, 93(6), 1923-1937.
6. Vitalis, A.; Pappu, R. V. J Comput Chem 2009, 30(5), 673-699.
7. Vitalis, A.; Wang, X.; Pappu, R. V. J Mol Biol 2008, 384(1), 279-297.
8. Vitalis, A.; Lyle, N.; Pappu, R. V. Biophys J 2009, *in press*.
9. Williamson, T. E.; Vitalis, A.; Crick, S. L.; Pappu, R. V. Nat Struct Mol Biol 2009, *to be submitted*.

10. La Spada, A. R.; Wilson, E. M.; Lubahn, D. B.; Harding, A. E.; Fischbeck, K. H. Nature 1991, 352(6330), 77-79.

11. MacDonald, M. E.; Ambrose, C. M.; Duyao, M. P.; Myers, R. H.; Lin, C.; Srinidhi, L.; Barnes, G.; Taylor, S. A.; James, M.; Groot, N.; MacFarlane, H.; Jenkins, B.; Anderson, M. A.; Wexler, N. S.; Gusella, J. F.; Bates, G. P.; Baxendale, S.; Hummerich, H.; Kirby, S.; North, M.; Youngman, S.; Mott, R.; Zehetner, G.; Sedlacek, Z.; Poustka, A.; Frischauf, A. M.; Lehrach, H.; Buckler, A. J.; Church, D.; Doucette-Stamm, L.; O'Donovan, M.; Riba-Ramirez, L.; Shah, M.; Stanton, V. P.; Strobel, S. A.; Draths, K. M.; Wales, J. L.; Dervan, P.; Housman, D. E.; Altherr, M.; Shiang, R.; Thompson, L. M.; Fielder, T.; Wasmuth, J. J.; Tagle, D.; Valdes, J.; Elmer, L.; Allard, M.; Castilla, L.; Swaroop, M.; Blanchard, K.; Collins, F. S.; Snell, R.; Holloway, T.; Gillespie, K.; Datson, N.; Shaw, D.; Harper, P. S. Cell 1993, 72(6), 971-983.

12. Orr, H. T.; Chung, M. Y.; Banfi, S.; Kwiatkowski Jr, T. J.; Servadio, A.; Beaudet, A. L.; McCall, A. E.; Duvick, L. A.; Ranum, L. P. W.; Zoghbi, H. Y. Nat Genet 1993, 4(3), 221-226.

13. Riley, B. E.; Orr, H. T. Genes Dev 2006, 20(16), 2183-2192.

14. Manto, M. U. The Cerebellum 2005, 4(1), 2-6.

15. Katsuno, M.; Adachi, H.; Waza, M.; Banno, H.; Suzuki, K.; Tanaka, F.; Doyu, M.; Sobue, G. Exp Neurol 2006, 200(1), 8-18.

16. Walker, F. O. Lancet 2007, 369(9557), 218-228.

17. Mangiarini, L.; Sathasivam, K.; Seller, M.; Cozens, B.; Harper, A.; Hetherington, C.; Lawton, M.; Trottier, Y.; Lehrach, H.; Davies, S. W.; Bates, G. P. Cell 1996, 87(3), 493-506.

18. Scherzinger, E.; Sittler, A.; Schweiger, K.; Heiser, V.; Lurz, R.; Hasenbank, R.; Bates, G. P.; Lehrach, H.; Wanker, E. E. Proc Natl Acad Sci U S A 1999, 96(8), 4604-4609.

19. Chen, S. M.; Ferrone, F. A.; Wetzel, R. Proc Natl Acad Sci U S A 2002, 99(18), 11884-11889.

20. Wanker, E. E. Biological Chemistry 2000, 381(9-10), 937-942.

21. Lipinski, M. M.; Yuan, J. Curr Opin Pharmacol 2004, 4(1), 85-90.

22. Hardy, J. A.; Higgins, G. A. Science 1992, 256(5054), 184-185.

23. Truant, R.; Atwal, R. S.; Desmond, C.; Munsie, L.; Tran, T. FEBS Journal 2008, 275(17), 4252-4262.

24. Ross, C. A.; Margolis, R. L.; Rosenblatt, A.; Ranen, N. G.; Becher, M. W.; Aylward, E. Medicine 1997, 76(5), 305-338.

25. Farrer, L. A. Am J Med Genet 1986, 24(2), 305-311.

26. Finsterer, J. Eur J Neurol 2009, 16(5), 556-561.

27. Jordan, C. L.; Lieberman, A. P. Curr Opin Pharmacol 2008, 8(6), 752-758.

28. Katsuno, M.; Adachi, H.; Kume, A.; Li, M.; Nakagomi, Y.; Niwa, H.; Sang, C.; Kobayashi, Y.; Doyu, M.; Sobue, G. Neuron 2002, 35(5), 843-854.

29. Nakamura, K.; Yoshida, K.; Miyazaki, D.; Morita, H.; Ikeda, S.-i. J Neurol Sci 2009, 278(1-2), 107.

30. Craufurd, D.; Snowden, J. In Huntington's disease; Bates, G.; Harper, P.; Jones, L., Eds.; Oxford University Press: New York, Oxford, 2002, p 62-94.

31. Brouwer, J. R.; Willemsen, R.; Oostra, B. A. Bioessays 2009, 31(1), 71-83.

32. Kennedy, L.; Evans, E.; Chen, C. M.; Craven, L.; Detloff, P. J.; Ennis, M.; Shelbourne, P. F. Hum Mol Genet 2003, 12(24), 3359-3367.

33. Kennedy, L.; Shelbourne, P. F. Hum Mol Genet 2000, 9(17), 2539.

34. Gomes-Peirera, M.; Foiry, L.; Gourdon, G. In Genetic instabilities and neurological disease; Wells, R. D.; Ashizawa, T., Eds.; Academic Press: Burlington, 2006, p 563-583.

35. Veitch, N. J.; Ennis, M.; McAbney, J. P.; Shelbourne, P. F.; Monckton, D. G. DNA Repair 2007, 6(6), 789-796.

36. Lloret, A.; Dragileva, E.; Teed, A.; Espinola, J.; Fossale, E.; Gillis, T.; Lopez, E.; Myers, R. H.; MacDonald, M. E.; Wheeler, V. C. Hum Mol Genet 2006, 15(12), 2015-2024.

37. Gonitel, R.; Moffitt, H.; Sathasivam, K.; Woodman, B.; Detloff, P. J.; Faull, R. L. M.; Bates, G. P. Proc Natl Acad Sci U S A 2008, 105(9), 3467-3472.

38. Mirkin, S. M. Curr Opin Struct Biol 2006, 16(3), 351-358.

39. Burnett, B.; Li, F.; Pittman, R. N. Hum Mol Genet 2003, 12(23), 3195-3205.

40. Zhuchenko, O.; Bailey, J.; Bonnen, P.; Ashizawa, T.; Stockton, D. W.; Amos, C.; Dobyns, W. B.; Subramony, S. H.; Zoghbi, H. Y.; Lee, C. C. Nat Genet 1997, 15(1), 62-69.

41. Helmlinger, D.; Hardy, S.; Eberlin, A.; Devys, D.; Tora, L. In Biochemical Society Symposium, 2006, p 155-163.

42. Stevanin, G.; Brice, A. The Cerebellum 2008, 7(2), 170-178.

43. Palazzolo, I.; Gliozzi, A.; Rusmini, P.; Sau, D.; Crippa, V.; Simonini, F.; Onesto, E.; Bolzoni, E.; Poletti, A. J Steroid Biochem Mol Biol 2008, 108(3-5), 245-253.

44. Zhang, C. L.; Zou, Y.; Yu, R. T.; Gage, F. H.; Evans, R. M. Genes Dev 2006, 20(10), 1308-1320.

45. Hubbard, T. J. P.; Aken, B. L.; Ayling, S.; Ballester, B.; Beal, K.; Bragin, E.; Brent, S.; Chen, Y.; Clapham, P.; Clarke, L.; Coates, G.; Fairley, S.; Fitzgerald, S.; Fernandez-Banet, J.; Gordon, L.; Graf, S.; Haider, S.; Hammond, M.; Holland, R.; Howe, K.;

Jenkinson, A.; Johnson, N.; Kahari, A.; Keefe, D.; Keenan, S.; Kinsella, R.; Kokocinski, F.; Kulesha, E.; Lawson, D.; Longden, I.; Megy, K.; Meidl, P.; Overduin, B.; Parker, A.; Pritchard, B.; Rios, D.; Schuster, M.; Slater, G.; Smedley, D.; Spooner, W.; Spudich, G.; Trevanion, S.; Vilella, A.; Vogel, J.; White, S.; Wilder, S.; Zadissa, A.; Birney, E.; Cunningham, F.; Curwen, V.; Durbin, R.; Fernandez-Suarez, X. M.; Herrero, J.; Kasprzyk, A.; Proctor, G.; Smith, J.; Searle, S.; Flicek, P. Nucleic Acids Res 2009, 37, D690-D697.

46. Katsuno, M.; Banno, H.; Suzuki, K.; Takeuchi, Y.; Kawashima, M.; Tanaka, F.; Adachi, H.; Sobue, G. Curr Mol Med 2008, 8(3), 221-234.

47. Li, L. B.; Yu, Z.; Teng, X.; Bonini, N. M. Nature 2008, 453(7198), 1107-1111.

48. Lim, J.; Crespo-Barreto, J.; Jafar-Nejad, P.; Bowman, A. B.; Richman, R.; Hill, D. E.; Orr, H. T.; Zoghbi, H. Y. Nature 2008, 452(7188), 713-718.

49. Lam, Y. C.; Bowman, A. B.; Jafar-Nejad, P.; Lim, J.; Richman, R.; Fryer, J. D.; Hyun, E. D.; Duvick, L. A.; Orr, H. T.; Botas, J.; Zoghbi, H. Y. Cell 2006, 127(7), 1335-1347.

50. Friedman, M. J.; Shah, A. G.; Fang, Z. H.; Ward, E. G.; Warren, S. T.; Li, S.; Li, X. J. Nat Neurosci 2007, 10(12), 1519-1528.

51. Schaffar, G.; Breuer, P.; Boteva, R.; Behrends, C.; Tzvetkov, N.; Strippel, N.; Sakahira, H.; Siegers, K.; Hayer-Hartl, M.; Hartl, F. U. Mol Cell 2004, 15(1), 95-105.

52. Friedman, M. J.; Wang, C. E.; Li, X. J.; Li, S. J Biol Chem 2008, 283(13), 8283-8290.

53. Cowan, K. J.; Diamond, M. I.; Welch, W. J. Hum Mol Genet 2003, 12(12), 1377-1391.

54. Takahashi, T.; Kikuchi, S.; Katada, S.; Nagai, Y.; Nishizawa, M.; Onodera, O. Hum Mol Genet 2008, 17(3), 345-356.

55. Wong, S. L. A.; Wing, M. C.; Chan, H. Y. E. FASEB J 2008, 22(9), 3348-3357.

56. Kayed, R.; Head, E.; Thompson, J. L.; McIntire, T. M.; Milton, S. C.; Cotman, C. W.; Glabel, C. G. Science 2003, 300(5618), 486-489.

57. Arrasate, M.; Mitra, S.; Schweitzer, E. S.; Segal, M. R.; Finkbeiner, S. Nature 2004, 431(7010), 805-810.

58. Sanchez, I.; Mahlke, C.; Yuan, J. Y. Nature 2003, 421(6921), 373-379.

59. Matsumoto, G.; Kim, S.; Morimoto, R. I. J Biol Chem 2006, 281(7), 4477-4485.

60. Warrick, J. M.; Chan, H. Y. E.; Gray-Board, G. L.; Chai, Y. H.; Paulson, H. L.; Bonini, N. M. Nat Genet 1999, 23(4), 425-428.

61. Kim, S.; Nollen, E. A. A.; Kitagawa, K.; Bindokas, V. P.; Morimoto, R. I. Nat Cell Biol 2002, 4(10), 826-831.

62. Muchowski, P. J.; Wacker, J. L. Nature Reviews Neuroscience 2005, 6(1), 11-22.

63. Choi, J. Y.; Ryu, J. H.; Kim, H. S.; Park, S. G.; Bae, K. H.; Kang, S.; Myung, P. K.; Cho, S.; Park, B. C.; Lee, D. H. Mol Cell Neurosci 2007, 34(1), 69-79.

64. Novoselova, T. V.; Margulis, B. A.; Novoselov, S. S.; Sapozhnikov, A. M.; Van Der Spuy, J.; Cheetham, M. E.; Guzhova, I. V. J Neurochem 2005, 94(3), 597-606.

65. Herbst, M.; Wanker, E. E. Neurodegener Dis 2007, 4(2-3), 254-260.

66. Parfitt, D. A.; Michael, G. J.; Vermeulen, E. G. M.; Prodromou, N. V.; Webb, T. R.; Gallo, J. M.; Cheetham, M. E.; Nicoll, W. S.; Blatch, G. L.; Chapple, J. P. Hum Mol Genet 2009, 18(9), 1556-1565.

67. Williams, A. J.; Knutson, T. M.; Colomer Gould, V. F.; Paulson, H. L. Neurobiol Dis 2009, 33(3), 342-353.

68. Nishikori, S.; Yamanaka, K.; Sakurai, T.; Esaki, M.; Ogura, T. Genes Cells 2008, 13(8), 827-838.

69. Kitamura, A.; Kubota, H.; Pack, C. G.; Matsumoto, G.; Hirayama, S.; Takahashi, Y.; Kimura, H.; Kinjo, M.; Morimoto, R. I.; Nagata, K. Nat Cell Biol 2006, 8(10), 1163-1170.

70. Tam, S.; Geller, R.; Spiess, C.; Frydman, J. Nat Cell Biol 2006, 8(10), 1155-1162.

71. Raychaudhuri, S.; Sinha, M.; Mukhopadhyay, D.; Bhattacharyya, N. P. Hum Mol Genet 2008, 17(2), 240-255.

72. Park, Y.; Hong, S.; Kim, S. J.; Kang, S. Mol Cells 2005, 19(1), 23-30.

73. Valera, A. G.; Díaz-Hernández, M.; Hernández, F.; Lucas, J. J. Brain Res Bull 2007, 72(2-3), 121-123.

74. Venkatraman, P.; Wetzel, R.; Tanaka, M.; Nukina, N.; Goldberg, A. L. Mol Cell 2004, 14(1), 95-104.

75. Wang, H.; Lim, P. J.; Karbowski, M.; Monteiro, M. J. Hum Mol Genet 2009, 18(4), 737-752.

76. Miyata, R.; Hayashi, M.; Tanuma, N.; Shioda, K.; Fukatsu, R.; Mizutani, S. J Neurol Sci 2008, 264(1-2), 133-139.

77. Sadri-Vakili, G.; Cha, J. H. J. Nat Clin Pract Neurol 2006, 2(6), 330-338.

78. Tanaka, M.; Machida, Y.; Nukina, N. J Mol Med 2005, 83(5), 343-352.

79. Ehrnhoefer, D. E.; Duennwald, M.; Markovic, P.; Wacker, J. L.; Engemann, S.; Roark, M.; Legleiter, J.; Marsh, J. L.; Thompson, L. M.; Lindquist, S.; Muchowski, P. J.; Wanker, E. E. Hum Mol Genet 2006, 15(18), 2743-2751.

80. Ignatova, Z.; Gierasch, L. M. Proc Natl Acad Sci U S A 2006, 103(36), 13357-13361.

81. Frid, P.; Anisimov, S. V.; Popovic, N. Brain Res Rev 2007, 53(1), 135-160.

82. Heiser, V.; Scherzinger, E.; Boeddrich, A.; Nordhoff, E.; Lurz, R.; Schugardt, N.; Lehrach, H.; Wanker, E. E. Proc Natl Acad Sci U S A 2000, 97(12), 6739-6744.

83. Popiel, H. A.; Nagai, Y.; Fujikake, N.; Toda, T. Neurosci Lett 2009, 449(2), 87-92.

84. Nagai, Y.; Tucker, T.; Ren, H.; Kenan, D. J.; Henderson, B. S.; Keene, J. D.; Strittmatter, W. J.; Burke, J. R. J Biol Chem 2000, 275(14), 10437.

85. Walsh, R.; Storey, E.; Stefani, D.; Kelly, L.; Turnbull, V. Neurotox Res 2005, 7(1-2), 43-57.

86. 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.; Richards, W. G. Neurobiol Dis 2002, 11(1), 111-122.

87. Lunkes, A.; Lindenberg, K. S.; Ben-Haem, L.; Weber, C.; Devys, D.; Landwehrmeyer, G. B.; Mandel, J. L.; Trottier, Y. Mol Cell 2002, 10(2), 259-269.

88. 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.; Ellerby, L. M. J Biol Chem 2007, 282(41), 30150-30160.

89. Kubodera, T.; Yokota, T.; Ohwada, K.; Ishikawa, K.; Miura, H.; Matsuoka, T.; Mizusawa, H. Neurosci Lett 2003, 341(1), 74-78.

90. Gafni, J.; Hermel, E.; Young, J. E.; Wellington, C. L.; Hayden, M. R.; Ellerby, L. M. J Biol Chem 2004, 279(19), 20211-20220.

91. Haacke, A.; Broadley, S. A.; Boteva, R.; Tzvetkov, N.; Hartl, F. U.; Breuer, P. Hum Mol Genet 2006, 15(4), 555-568.

92. Chen, S.; Wetzel, R. Protein Sci 2001, 10(4), 887-891.

93. Ferrone, F. In Amyloid, Prions, And Other Protein Aggregates, 1999, p 256-274.

94. Zhao, D.; Moore, J. S. Org Biomol Chem 2003, 1(20), 3471-3491.

95. Bhattacharyya, A. M.; Thakur, A. K.; Wetzel, R. Proc Natl Acad Sci U S A 2005, 102(43), 15400-15405.

96. Oosawa, F.; Kasai, M. J Mol Biol 1962, 4, 10-21.

97. Bernacki, J. P.; Murphy, R. M. Biophys J 2009, 96(7), 2871-2887.

98. Chen, S.; Berthelier, V.; Yang, W.; Wetzel, R. J Mol Biol 2001, 311(1), 173-182.

99. Lee, C. C.; Walters, R. H.; Murphy, R. M. Biochemistry 2007, 46(44), 12810-12820.

100. Poirier, M. A.; Li, H. L.; Macosko, J.; Cai, S. W.; Amzel, M.; Ross, C. A. J Biol Chem 2002, 277(43), 41032-41037.

101. Tanaka, M.; Morishima, I.; Akagi, T.; Hashikawa, T.; Nukina, N. J Biol Chem 2001, 276(48), 45470-45475.

102. Ignatova, Z.; Thakur, A. K.; Wetzel, R.; Gierasch, L. M. J Biol Chem 2007, 282(50), 36736-36743.

103. Pappu, R. V.; Wang, X.; Vitalis, A.; Crick, S. L. Arch Biochem Biophys 2007, 469(1), 132-141.

104. 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.; Wetzel, R. Nat Struct Mol Biol 2009, 16(4), 380-389.

105. Temussi, P. A.; Masino, L.; Pastore, A. EMBO J 2003, 22(3), 355-361.

106. Dima, R. I.; Thirumalai, D. Protein Sci 2002, 11(5), 1036-1049.

107. Nguyen, H. D.; Hall, C. K. Proc Natl Acad Sci U S A 2004, 101(46), 16180-16185.

108. Nguyen, H. D.; Hall, C. K. J Am Chem Soc 2006, 128(6), 1890-1901.

109. Pellarin, R.; Caflisch, A. J Mol Biol 2006, 360(4), 882-892.

110. Bennett, M. J.; Huey-Tubman, K. E.; Herr, A. B.; West, A. P.; Ross, S. A.; Bjorkman, P. J. Proc Natl Acad Sci U S A 2002, 99(18), 11634-11639.

111. Masino, L.; Kelly, G.; Leonard, K.; Trottier, Y.; Pastore, A. FEBS Lett 2002, 513(2-3), 267-272.

112. Robertson, A. L.; Horne, J.; Ellisdon, A. M.; Thomas, B.; Scanlon, M. J.; Bottomley, S. P. Biophys J 2008, 95(12), 5922-5930.

113. Wang, X. L.; Vitalis, A.; Wyczalkowski, M. A.; Pappu, R. V. Prot Struct Funct Bioinf 2006, 63(2), 297-311.

114. Crick, S. L.; Jayaraman, M.; Frieden, C.; Wetzel, R.; Pappu, R. V. Proc Natl Acad Sci U S A 2006, 103(45), 16764-16769.

115. Krull, L. H.; Wall, J. S. Biochemistry 1966, 5(5), 1521-1527.

116. Tran, H. T.; Pappu, R. V. Biophys J 2006, 91(5), 1868-1886.

117. Singh, V. R.; Lapidus, L. J. J Phys Chem B 2008, 112(42), 13172-13176.

118. Li, P.; Huey-Tubman, K. E.; Gao, T.; Li, X.; West Jr, A. P.; Bennett, M. J.; Bjorkman, P. J. Nat Struct Mol Biol 2007, 14(5), 381-387.

119. Scherzinger, E.; Lurz, R.; Turmaine, M.; Mangiarini, L.; Hollenbach, B.; Hasenbank, R.; Bates, G. P.; Davies, S. W.; Lehrach, H.; Wanker, E. E. Cell 1997, 90(3), 549-558.

120. Chen, S. M.; Berthelier, V.; Hamilton, J. B.; O'Nuallain, B.; Wetzel, R. Biochemistry 2002, 41(23), 7391-7399.

121. Sipe, J. D.; Cohen, A. S. J Struct Biol 2000, 130(2-3), 88-98.

122. Perutz, M. F.; Johnson, T.; Suzuki, M.; Finch, J. T. Proc Natl Acad Sci U S A 1994, 91(12), 5355-5358.

123. Perutz, M. F.; Finch, J. T.; Berriman, J.; Lesk, A. Proc Natl Acad Sci U S A 2002, 99(8), 5591-5595.

124. Sikorski, P.; Atkins, E. Biomacromolecules 2005, 6(1), 425-432.

125. Sharma, D.; Shinchuk, L. M.; Inouye, H.; Wetzel, R.; Kirschner, D. A. Prot Struct Funct Bioinf 2005, 61(2), 398-411.

126. Armen, R. S.; Bernard, B. M.; Day, R.; Alonso, D. O. V.; Daggett, V. Proc Natl Acad Sci U S A 2005, 102(38), 13433-13438.

127. Zanuy, D.; Gunasekaran, K.; Lesk, A. M.; Nussinov, R. J Mol Biol 2006, 358(1), 330-345.

128. Marchut, A. J.; Hall, C. K. Prot Struct Funct Bioinf 2007, 66(1), 96-109.

129. Merlino, A.; Esposito, L.; Vitagliano, L. Prot Struct Funct Bioinf 2006, 63(4), 918-927.

130. Esposito, L.; Paladino, A.; Pedone, C.; Vitagliano, L. Biophys J 2008, 94(10), 4031-4040.

131. Ogawa, H.; Nakano, M.; Watanabe, H.; Starikov, E. B.; Rothstein, S. M.; Tanaka, S. Comput Biol Chem 2008, 32(2), 102-110.

132. Nagai, Y.; Inui, T.; Popiel, H. A.; Fujikake, N.; Hasegawa, K.; Urade, Y.; Goto, Y.; Naiki, H.; Toda, T. Nat Struct Mol Biol 2007, 14(4), 332-340.

133. 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. Cell Cycle 2007, 6(23), 2970-2981.

134. 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.; Ross, C. A. J Biol Chem 2009, 284(16), 10855-10867.

135. Orr, A. L.; Li, S.; Wang, C. E.; Li, H.; Wang, J.; Rong, J.; Xu, X.; Mastroberardino, P. G.; Greenamyre, J. T.; Li, X. J. Journal of Neuroscience 2008, 28(11), 2783.

136. Rockabrand, E.; Slepko, N.; Pantalone, A.; Nukala, V. N.; Kazantsev, A.; Marsh, J. L.; Sullivan, P. G.; Steffan, J. S.; Sensi, S. L.; Thompson, L. M. Hum Mol Genet 2007, 16(1), 61-77.

137. Darnell, G.; Orgel, J. P. R. O.; Pahl, R.; Meredith, S. C. J Mol Biol 2007, 374(3), 688-704.

138. Bhattacharyya, A.; Thakur, A. K.; Chellgren, V. M.; Thiagarajan, G.; Williams, A. D.; Chellgren, B. W.; Creamer, T. P.; Wetzel, R. J Mol Biol 2006, 355(3), 524-535.

CHAPTER II. QUANTIFICATION OF CONFORMATIONAL EQUILIBRIA OF MONOMERIC POLYGLUTAMINE: INSIGHTS FROM ANALYSIS BASED ON POLYMER PHYSICS

II.1 Preamble

The project of using a brute-force computational approach to quantify conformational equilibria of monomeric polyglutamine employing an explicit representation of solvent has a history which extends beyond the manuscript[1] incorporated into this chapter. Previous work[2] had revealed interesting and somewhat similar insights for peptides of length five and fifteen but the analysis then had focused much more on topology and hydrogen bond statistics. Part of the reason is that the data were too immature to arrive at conclusions of similar rigor as shown below. One of the focal points of this chapter is the interdisciplinary approach: we apply analysis concepts borrowed from polymer physics to a biological polymer at atomic resolution (see I.1). Another focal point is the quantification of the difficulty of sampling a system of this complexity by establishing timescales for conformational re-arrangement. The major conclusion with respect to the latter is that that difficulty becomes overwhelming very quickly (see II.5), hence motivating the development of an alternative sampling technique as detailed in Chapter III.

As mentioned above, this chapter is based on a research article which appeared in the Biophysical Journal in 2007.[1] Xiaoling Wang is a co-author on this manuscript and hence her contributions need to be established in detail: she

ran and partially analyzed an original set of calculations for Acetyl-Q_{20}-N-methylamide (Q_{20}). The scope of that analysis was slightly different from what is presented here. The calculations themselves turned out to be insufficient and all results presented in this chapter are based on an independent set of data not generated or analyzed by her. She therefore established the groundwork for this project as is also evidenced by our prior work.[2]

II.2 Introduction to the Application of Polymer Physics on Conformational Equilibria of IDPs

Intrinsically disordered proteins (IDPs) are functional proteins that do not fold into well-defined, ordered tertiary structures under physiological conditions.[3-6] These proteins are termed intrinsically disordered because disorder prevails under non-denaturing conditions and amino acid sequence encodes the propensity to be disordered. Generic IDP sequences have a combination of low overall hydrophobicity[7] and low sequence complexity.[8] Consequently, they include low complexity polymers composed predominantly of polar amino acids such as glutamine which is found frequently in aggregation-prone sequences.[9] The question of how disorder is used in function – whether it is related to its native biological role or to disease pathology – will remain unanswered pending the availability of accurate physical models for conformational equilibria of IDPs.[6] Conformational equilibria refer to ensemble averages and spontaneous fluctuations of structural properties of IDPs in their native milieus.

Theories based on the physics of polymer solutions are relevant for

describing conformational equilibria of IDPs.[10] The focus in these theories is on global measures such as the ensemble-averaged radius of gyration, $\langle R_g \rangle$.[11] The balance between chain-chain and chain-solvent interactions is determined by the nature of solvent milieus, which are classified as being good or poor solvents.[12,13] The scaling of $\langle R_g \rangle$ with chain length N is written as $\langle R_g \rangle = R_o N^{\nu}$. In a good solvent, the main repeating unit is chemically equivalent to the surrounding solvent, the effective chain-chain interactions are strictly repulsive, and $\langle R_g \rangle \sim N^{0.59}$. In a poor solvent, attractive interactions dominate and the result is a preference for an ensemble of compact conformations such that $\langle R_g \rangle \sim N^{0.33}$.[14] In the simplest of polymer frameworks, conformational ensembles for IDPs in aqueous milieus can be classified either as disordered swollen coils in a good solvent or compact, albeit disordered globules in a poor solvent. Which of these classifications best suits the description of conformational ensembles for the disease-related IDP polylgutamine in water? This is one of the core questions of this chapter. Specifically:

1. *Is it possible to make quantitative assessments regarding the quality of a solvent milieu for polyglutamine at a single chain length using data obtained from molecular simulations*? To answer this question, we study polyglutamine of chain length N=20. Specifically, we compared results from analysis of multiple replica molecular dynamics (MRMD) for Q_{20} in water to data from two sets of Metropolis Monte Carlo simulations for reference ensembles in good and poor solvents. The Monte Carlo simulations employed here are routinely

used in the polymer physics literature and are based on the use of generic Hamiltonians that lack the specificities of chain-chain and chain-solvent interactions.[15,16] The comparative analysis is guided by the use of polymer theories,[17,18] which make specific predictions regarding variations of order parameters such as the scaling of internal distances, angular correlation functions, and radial density profiles as a function of solvent quality. We show that the comparative analysis leads unequivocally to the identification that water is a poor solvent for Q_{20}, which is consistent with experimental results obtained using fluorescence correlation spectroscopy (FCS).[19] The main highlight of this analysis is that it can be adapted to classify the nature of disorder for any IDP sequence, in particular those with low complexity.[8]

2. *Why is water a poor solvent for polyglutamine*? The observation that water is a poor solvent for polyglutamine can be inferred from its strong aggregation propensity.[20-22] However, it seems counterintuitive that a system composed entirely of polar moieties readily forms aggregates given that the building blocks of polyglutamine, *i.e.*, primary and secondary amides, are freely miscible with water.[23,24] If anything, the high miscibility of model compounds suggests that water should be a good solvent for polyglutamine. Obviously, the concatenation into a polymer alters the solvation properties of amide groups. Here, we present a preliminary analysis based on comparisons of data from simulations of aqueous solutions of amide mixtures to that of Q_{20} in water. Based on this analysis, we propose that favorable intra-backbone interactions

in the polymer provide at least part of the driving force for the collapse of polyglutamine in water.

3. *What is the nature of conformational relaxation dynamics for polyglutamine?* Polyglutamine forms aggregates, albeit very slowly.[25] Chuang *et al.*[26] have proposed that the rate limiting step for aggregation of polymers in poor solvents is conformational relaxation within polymer globules. Consistent with this prediction, we find that – although the collapse transition for Q_{20} in water is rapid (ca. 5ns) – the time scales for conversion between distinct compact conformations are very slow and the dynamics are akin to structural relaxation in glassy systems.[27] We also show that the glassy behavior of Q_{20} in water is uncovered using the MRMD methodology employed in our work.

The rest of this chapter is structured as follows: A thorough presentation of the methods (II.3) is followed by the results (II.4). We conclude by placing the latter in the broader context of the field: specifically the scope of this thesis (II.5).

II.3 Simulation Details and Methods of Analysis

II.3.1. Potential Functions for Simulating Conformational Equilibria of Polymeric Reference States

Reference conformational equilibria of disordered polymers in good and poor solvents can be simulated using generic, implicit solvent models.[15,16] In this approach,[28] conformational equilibria for chains in good solvents are simulated using interatomic interactions based on a purely repulsive, inverse power potential as shown in Equation 2-1:

$$U_{\mathrm{EV}} = 4\sum_{i}\sum_{j<i}\varepsilon_{ij}\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} \qquad (2\text{-}1)$$

Equation 2-1 corresponds to the so-called excluded volume (EV) limit, wherein only steric interactions are included. Simulations of conformational equilibria in the EV limit provide a good mimic for equilibria in good solvents. Conversely, the non-specific drive of a chain to sequester itself from making contacts with a poor solvent can be captured by adding attractive van der Waals interactions to the repulsive potentials from the EV limit. This model, based on the Lennard-Jones functional form, is shown in Equation 2-2 and is termed the LJ model.

$$U_{\mathrm{LJ}} = 4\sum_{i}\sum_{j<i}\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \qquad (2\text{-}2)$$

In Equations 2-1 and 2-2, r_{ij} denote distances between any two non-bonded atoms, σ_{ij} are contact distances, and ε_{ij} are the LJ dispersion parameters. For the EV limit, Equation 2-1, the parameters for σ_{ij} and ε_{ij} are those used in previous work by Tran and Pappu.[29] Conversely, for the LJ model, Equation 2-2, we used the parameters from the OPLS-AA/L force field.[30] Here, standard geometric combination rules were used to obtain the σ_{ij} and ε_{ij} from the σ_{ii} and ε_{ii}. These choices are justified on the following grounds: The σ_{ij} values used in previous work were derived from Pauling's parameterization, which in turn reproduce data for the heats of fusion of model compounds. These σ_{ij} values can be used in purely repulsive potentials and it has been shown that in conjunction with the ε_{ij} derived from atomic polarizabilities these parameters allow the

reproduction of accurate Ramachandran maps.[29] Conversely, the values of σ_{ij} in the OPLS-AA/L force field are co-parameterized with ε_{ij} to reproduce the heats of vaporization and densities of neat liquids. Therefore, the σ_{ij} values in OPLS-AA/L are too large to be used in purely repulsive potentials. However, use of the OPLS-AA/L parameters for the LJ model guarantees that the densities of the maximally compact reference globules are similar to those expected for globules populated by chains in explicit water.

II.3.2. Simulations of Reference Conformational Equilibria

We carried out Metropolis Monte Carlo simulations, as described in previous work,[28,29] to simulate reference conformational equilibria for polyglutamine peptides using the EV and LJ models. In these simulations, the degrees of freedom were the backbone and sidechain dihedral angles of an isolated chain (compare Chapters III-VI). We obtained two sets of data using each of the models shown in Equations 2-1 and 2-2. In the first set, we carried out simulations for a series of chain lengths to demonstrate that ensemble averaged radii of gyration ($\langle R_g \rangle$) scale with chain length as ca. $N^{0.6}$ for the EV model and as $N^{0.33}$ for the LJ-model. The scaling of $\langle R_g \rangle$ with chain length N was obtained by gathering statistics for peptides of the form: Ace-(Gln)$_N$-Nme, where Ace denotes the acetyl group and Nme stands for N-methylamide. For the EV limit, N=50, 75, 100, 150, and 250 and for the LJ model, we simulated equilibria for N=24, 27, 33, 36, 40, 47. The simulation temperatures were T=298K and T=425K for the EV and LJ models, respectively. We used a higher temperature

in simulations based on the LJ model to improve the efficiency with which conformational space is sampled and to reduce the error bars in our estimates for polymeric properties. Given the high melting temperature for the LJ model, at T=298K we would have needed simulations that were orders of magnitude longer in order to obtain converged estimates; hence 425K was chosen as the simulation temperature.

As noted above, the purpose of the Monte Carlo simulations was to demonstrate that the two models, *viz.* EV and LJ, reproduce the scaling behaviors for polymers in good and poor solvents, respectively. The EV limit calculations were carried out for longer chains to overcome the finite size artifacts because the thickness of the polymer "tube" has to be negligible when compared to its contour length. For polyglutamine, this requirement does not hold true for chains with N<50. In contrast, finite size effects play a minor role for quantifying the scaling law for chains in a poor solvent. This is true so long as N is larger than the length of locally stiff segments, approximately seven residues.[28] The chain lengths used for calculations in the globular limit were therefore chosen in correspondence with recent FCS studies.[19] In addition to the simulations used to quantify scaling laws, we also simulated conformational equilibria for Q_{20} using both the EV and LJ models. As is shown in II.4, the comparative analysis between ensembles obtained for Q_{20} using the EV, LJ, and molecular mechanics potentials in explicit solvent allows us to assess if the conformational equilibria for Q_{20} in water are congruent with those of chains in poor versus good solvents.

II.3.3. Setup of Molecular Dynamics Simulations for Q_{20}

To characterize conformational equilibria in water we used an approach that we refer to as multiple replica molecular dynamics (MRMD). This approach relies on the use of data from a large number of independent simulations and the advantage is that data are gathered using multiple independent simulations as opposed to a single, long, and potentially uninformative simulation. Conformational space is explored more efficiently by relying on the underlying stochasticity of phase space trajectories given different initial positions and velocities.

We used version 3 of the GROMACS simulation package[31] for all MD simulations. In this work, we report data from MRMD simulations applied to the peptide Q_{20} in water at T=298K. We simulated 60 independent replicas. For the peptide we used the OPLS-AA/L force-field.[30] Peptides were soaked in a bath of 8952 TIP3P water molecules.[32] We randomly selected peptide conformations from the EV ensemble for this purpose. By adding or deleting water molecules, we ensured that we ended up with the same number density for all replicas. In each case, a steepest-descent minimization to remove steric clashes was followed by an equilibration run of 11ns in the isothermal-isobaric ensemble (T=298K, P=1bar). The final configuration of the latter was used as the starting point for the production run of 50ns length. Therefore, the total simulation time for each of the 60 independent simulations was 61ns for a cumulative simulation time of approximately 3.7μs.

The leap-frog integrator was used with a time-step of 2fs. The temperature was maintained through the Berendsen thermostat[33] with a coupling time of 0.2ps. Similarly, constant pressure was maintained by the Berendsen manostat[33] with a coupling time of 1ps and a compressibility of $4.5\times10^{-5}bar^{-1}$. It is important to point out that the Berendsen weak-coupling scheme does not rigorously achieve sampling of the canonical ensemble.[33] However, for the robust equilibrium properties assessed here, the impact of the quenching of energetic fluctuations can be expected to be minor. Additionally, integrator noise led to the actual simulation temperatures being slightly higher than the specified bath temperature (298K). Again, artifacts deriving from such minor deviations are expected to be insignificant for the range of properties studied here. The average size of the cubic box throughout the simulations was roughly 65.4Å with negligible volume fluctuations. Bond lengths for atoms in the polypeptide were constrained using the LINCS algorithm[34] and the rigidity of water molecules was achieved using the SETTLE algorithm.[35] For non-bonded interactions, we employed 10-14Å twin-range cutoffs. Both LJ and Coulomb interactions within distances of 10Å were calculated at every step. Conversely, interactions within the twin-range (10-14Å) were re-calculated every ten steps, as were neighbor lists. The reaction field (RF) method[36] was used as a correction term for polar interactions beyond 14Å. For each of the sixty independent simulations, structures of the peptide alone were saved once every 4ps for subsequent analysis.

II.3.4. Setup of Simulations for Aqueous Solutions of Model Compounds

To assess the differences between polyamides (such as polyglutamine) versus amides in water we carried out simulations of aqueous mixtures of amides. The systems studied were aqueous mixtures of *trans*-N-methylacetamide (NMA) and propionamide (PPA) in water; NMA is a model compound mimic of the peptide backbone (a secondary amide) whereas PPA is a mimic of the sidechain (a primary amide). We followed the simulation protocol described for Q_{20}. The amides were modeled using the OPLS-AA/L force field,[30] and we used the TIP3P model for water molecules. To achieve concentrations of 1 m, 2 m, and 4 m respectively, 15, 30, and 60 molecules of each amide were soaked in a box of 833 water molecules and equilibrated for mixing purposes for 1ns in the canonical (*NVT*) ensemble at *T*=298K. The production run was carried out in the isothermal-isobaric (*T*=298K, *P*=1bar) ensemble for 50ns after an extra equilibration period of 200ps. Ten such trajectories were run for each concentration and the snapshots of the amide configurations, which were saved every 10ps, were analyzed to calculate site-site pair correlation functions.

II.3.5. Reliability Analysis

Given n_s independent trajectories, the standard error in our data was estimated by computing the average of an observable for each trajectory. The standard error is defined as the standard deviation in n_s independent estimates of the mean. Our procedure for computing the standard error is an adaptation of conventional block averaging methods. The difference is that the size of the

block being averaged over is the length of an individual trajectory. The standard deviation of the trajectory-averaged structural quantities yields the standard error indicated by error bars in the plots. This approach for calculating error bars is reasonable because data from different trajectories are in fact truly uncorrelated.

II.3.6. Calculations of Intra-Polymer Site-Site Correlation Functions

Consider all unique pairs of backbone donor (N) and acceptor atoms (O), respectively. For generality, we shall use the labels *A* and *B* to refer to these atom pairs. Let $h_W(r_{AB})$ denote the histogram of interatomic distances obtained from analysis of MRMD simulation data for Q_{20} in water. Additionally, let $h_D(r_{AB})$ be the histogram obtained by gathering statistics from simulations based on an appropriate default model. Given the two histograms, $h_W(r_{AB})$ and $h_D(r_{AB})$, the desired site-site correlation function is defined as:

$$g_{AB}(r) = \frac{h_W(r_{AB})}{h_D(r_{AB})} \tag{2-3}$$

It is important to emphasize that the choice for the default model determines the profile we obtain for $g_{AB}(r)$. The standard non-interacting model used in the theory of liquids is the so-called ideal gas prior. In this model, the sites are parts of rigid molecules that are free to translate and rotate around each other. The applicability of this default model for polymers is questionable because the resultant profiles one obtains for $g_{AB}(r)$ are dominated by the presence of chain connectivity in the real chain, which increases the effective concentration of repeating units with respect to each other. Therefore, we

constructed intra-chain site-site correlation functions using a so-called ideal chain model, which is analogous to the freely rotating chain model of Flory.[12] In this model, bond lengths and bond angles are held fixed at equilibrium values[28] and the peptide unit is held fixed in the *trans*-configuration. An ensemble of freely-rotating chain conformations is generated by ignoring (turning off) all non-bonded interactions, including excluded volume effects. Histograms, $h_D(r_{AB})$, constructed using the resultant ensemble include the effects of chain connectivity and exclude the effects of intra-chain and chain-solvent interactions.

II.4 Results

II.4.1. Demonstration of the Validity of Reference Models

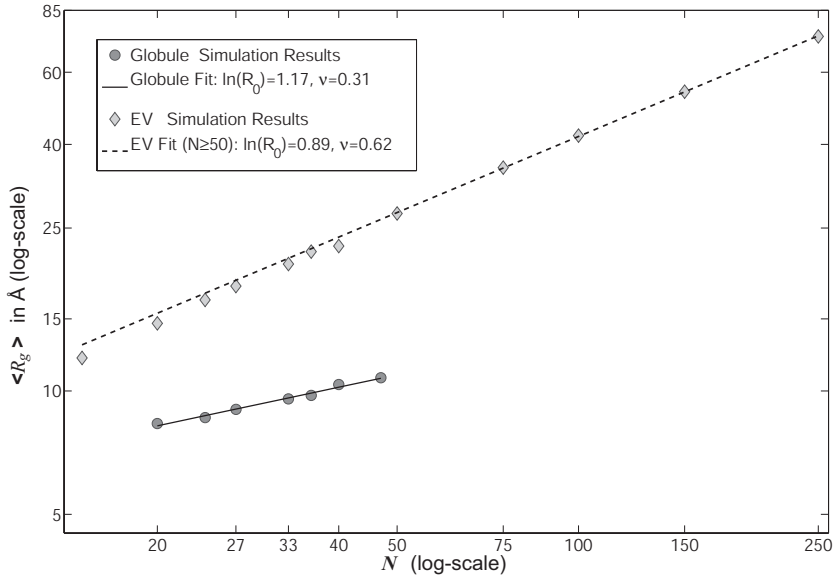

Figure 2.1: **Scaling laws for the two reference models** (see Equations 2-1 and 2-2). The fit for the EV limit is done only over the last five points. As can be seen, finite-size effects cause the data for shorter chain lengths to fall off this line. Including these points would significantly overestimate the scaling exponent. In the globular reference state,

finite-size effects are restricted to much shorter chain lengths. The theoretical exponent of ~0.33 is slightly underestimated.

Figure 2.1 shows the scaling of $\langle R_g \rangle$ versus chain length N for polyglutamine in the EV and LJ limits, respectively. In the log-log plots shown, the slopes provide an estimate of the scaling exponent. We find that slopes for polyglutamine in the EV and LJ limits are similar to the theoretical values of 0.59 and 0.33 in good and poor solvents, respectively. Deviations from theoretical values are primarily due to finite-size effects, *i.e.*, the fact that we did not gather data for very long chains. In properly converged simulations, the scaling exponent in the EV limit will be over-estimated when there are finite size artifacts. This is because short chains in the EV limit have a smaller, apparent $\langle R_g \rangle$ when compared to the theoretical prediction. Conversely, finite size artifacts lead to an underestimation of the poor solvent exponent. This is because short chains have a larger apparent $\langle R_g \rangle$, which is precisely what we find.

The preceding analysis demonstrates that conformational equilibria simulated using the EV and LJ models provide limiting distributions for disordered polypeptides in good versus poor solvents. Due to the extensive computational cost of the simulations in explicit water (see below) we cannot determine the scaling exponent, which requires very expensive simulations for multiple chain lengths. Instead, analyses of specific polymeric measures for Q_{20} in water were compared to those of Q_{20} in the EV and LJ limits, respectively. This allows us to make definitive conclusions regarding the solvent quality of water for polyglutamine.

II.4.2. Quantification of Polymeric Properties

After establishing the validity of the reference models with respect to their universal scaling behavior, we now turn our attention to a detailed analysis of polymeric properties of Q_{20} in aqueous solution. As outlined in II.2, the analysis relies on structural quantities inspired by polymer physics. This is crucial since in the absence of a canonical fold it is not as easy to define informative metrics (see below).

Comparative analysis of the distribution of shapes and sizes

For a specific conformation of a polymer, the shape and size are quantified using the gyration tensor defined as:

$$\mathbf{T} = \frac{1}{Z_m} \cdot \sum_{i=1}^{Z_m} \left(\mathbf{r}_i - \bar{\mathbf{r}}\right) \otimes \left(\mathbf{r}_i - \bar{\mathbf{r}}\right) \tag{2-4}$$

Here, Z_m is the number of atoms in the molecule, $\mathbf{r}_i$ are the position vectors of individual atoms, $\bar{\mathbf{r}}$ is the position vector of the centroid, and the symbol $\otimes$ refers to the dyadic product. If we use $\lambda_{1,2,3}$ to denote the eigenvalues of $\mathbf{T}$, the radius of gyration (R_g), the measure of size, and asphericity (δ), which measures chain shape are given as:

$$\begin{aligned} R_g &= \sqrt{\lambda_1 + \lambda_2 + \lambda_3} \\ \delta &= 1 - 3\left(\frac{\lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_1}{\left(\lambda_1 + \lambda_2 + \lambda_3\right)^2} \right) \end{aligned} \tag{2-5}$$

For a perfect sphere, δ=0, and for a perfect rod, δ=1; for intermediate values, the chain assumes ellipsoidal shapes. Therefore, δ quantifies the degree to which chain shape deviates from that of a perfect sphere. This measure of shape has been very useful for analyzing asymmetry in protein structures[37] and for the analysis of average shapes of denatured proteins.[28]

Two-dimensional histograms, *viz.* $\rho(R_g, \delta)$ in the space spanned by the two parameters R_g and δ, provide insights regarding the preferred shapes and sizes of a molecule.[2] Figure 2.2 shows these distributions for Q_{20} in water and for the two reference models:

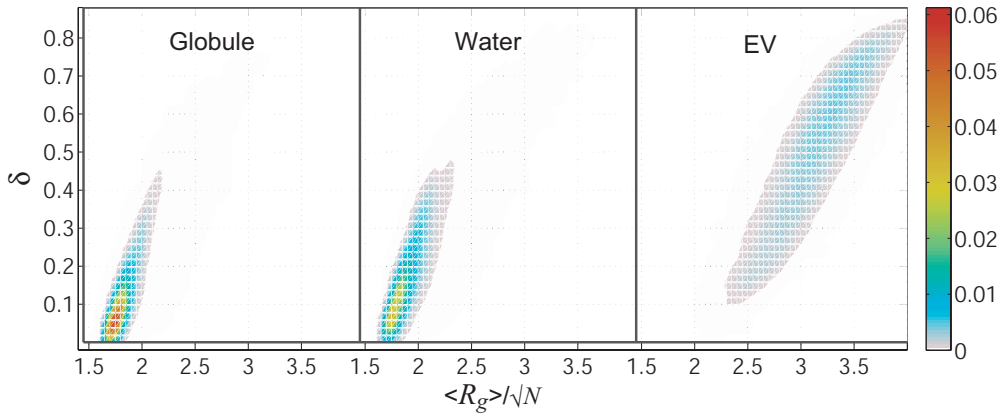

Figure 2.2: **Two-dimensional histograms of the normalized radius of gyration and asphericity** (see Equation 2-4) **for Q_{20} in water and the two reference models**. The data are binned with a spacing of 0.05Å on the R_g-axis and 0.02 on the δ-axis, respectively. For the purpose of clarity, the colors are slightly offset from the white background.

Conformations with low asphericity and low R_g are favored for Q_{20} in water. This is suggestive of water being a poor solvent for Q_{20}. This point is reinforced

by favorable comparison of histograms in water to those obtained for the globular reference ensemble using the LJ model. The only difference is that the latter are characterized by smaller-scale fluctuations. In stark contrast, the peptides in the EV limit prefer conformations with larger R_g and asphericity values. Even more importantly, there is no overlap between histograms obtained in the EV limit versus those for either Q_{20} in water or Q_{20} in the reference globule. Polymers of the requisite length have access to three distinct phases, *viz*. the globule, coil, and rod phases.[18] The data shown in Figure 2.2 support the conclusion that conformational equilibria for Q_{20} in water and calculated using the LJ model are consistent with the globule phase while equilibria in the EV limit are consistent with those of the coil phase.

Collapse does not mean order

One might be tempted to speculate that Q_{20} prefers a specific globular structure in water. If true, such an observation would be incongruent with experimental observations according to which soluble and monomeric polyglutamine peptides are described as being disordered by measures such as CD[20] or NMR.[38]

Figure 2.3 shows that our results are consistent with interpretations of experimental data. The inter-residue contact maps show no preference for specific contacts. We can, however, distinguish two classes of disorder: (i) disorder under the constraint of dense packing results in relatively large contact probabilities (see Panels B and C), and (ii) disorder in the swollen-coil state with

very low contact probabilities (see Panel A). The preferred contacts in the EV limit are exclusively local. Conversely, in both the LJ globule as well as in water, long-range contacts (sequence spacing >10) are actually more likely than mid-range contacts (sequence spacing 5-9). Local contacts are enhanced in the aqueous case vis-à-vis the LJ globule. We attribute these differences between the LJ globule and the aqueous globule to specific local interactions present in the latter,[2] a feature that is missing in the case of the LJ globule.

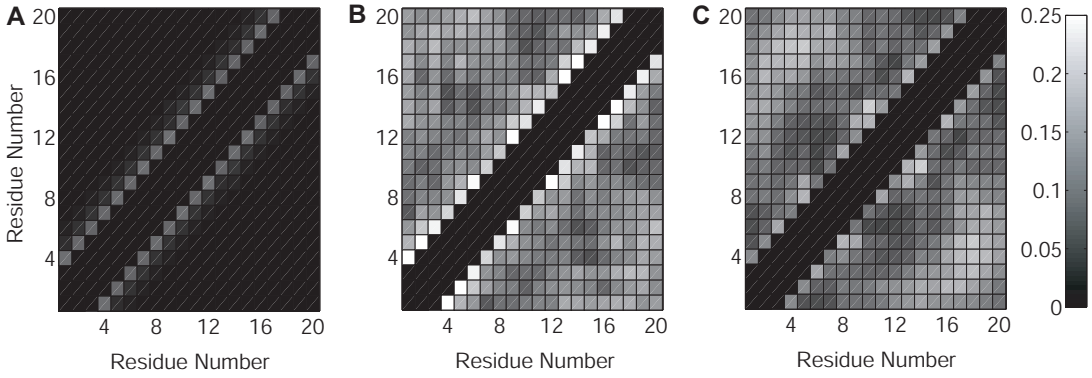

Figure 2.3: **Contact maps for Q_{20} in water** (Panel B)**, in the EV limit** (Panel A)**, and in the globular limit** (Panel C). Grayscale indicates the frequency of observing a given residue-residue contact throughout the simulation. Short-range contacts are excluded to enhance the signal-to-noise ratio. A contact is defined by any two atoms k and l from residues i and j having a distance less than 3Å. The maps are by definition symmetric.

One might argue that our analysis of disorder observed for Q_{20} in water masks the identification of secondary structure, since α-helices or β-sheets with highly variable registry might be possible. However, previous analysis of backbone segments confirmed that there is little to no stable canonical

secondary structure.[2] Similar conclusions were drawn from the current dataset (data not shown).

Scaling of internal distances with sequence separation

The first polymeric measure we quantify is the scaling of internal distances with sequence separation:

$$\left\langle R_{ij} \right\rangle = \left\langle \frac{1}{Z_{ij}} \cdot \sum_{m \in i} \sum_{n \in j} \left| \mathbf{r}_m^i - \mathbf{r}_n^j \right| \right\rangle \qquad (2\text{-}6)$$

In Equation 2-6, the $\mathbf{r}_m^i$ and $\mathbf{r}_n^j$ denote the position vectors of atoms m and n, which are part of residues i and j, respectively, and Z_{ij} is the number of unique pairwise distances between the two residues. As in all equations, the angular brackets indicate the average over all trajectories and all saved snapshots. Plotted as a function of sequence separation, it is expected that for chains in a good solvent $\langle R_{ij} \rangle \sim |j\text{-}i|^{0.59}$,[39] which is true in the EV limit.[28] In a good solvent, polymers behave like fractal objects, *i.e.*, internal distances scale with sequence separation the way end-to-end distances scale with chain length. Figure 2.4 shows that the scaling of internal distances in the EV limit ensemble agrees with the theoretical scaling law. Significant deviations occur at small sequence separations, for which the local rigidity and detailed structure of the polymer modulate the limiting behavior. Similar observations were made by Ding *et al.*[40] in their analysis of the scaling behavior of proteins near and above the folding transition.

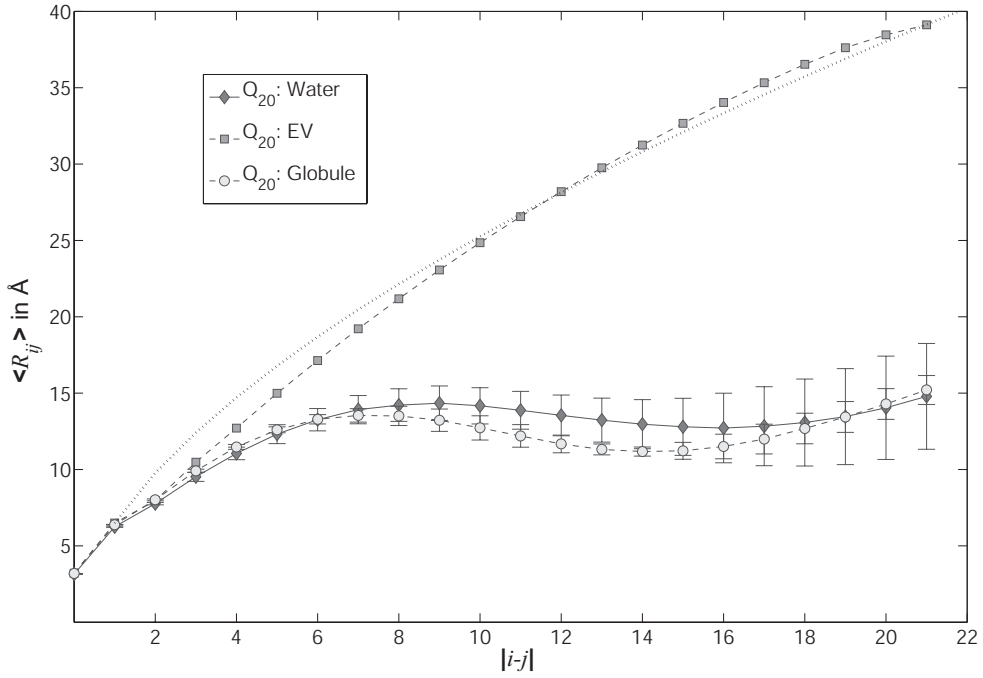

Figure 2.4: **The scaling of average internal distances as a function of sequence separation** (see Equation 2-6). A theoretical good solvent scaling law is indicated by the dotted line. Standard errors are indicated by error bars for the data in water and the globular reference state. Errors are negligible for the EV ensemble and hence not shown. The polypeptide caps are included in this analysis, which is why there are effectively 22 residues in the chain. Unless otherwise noted, lines are drawn exclusively to guide the eye. This is true for all subsequent figures in this thesis.

Conversely, for chains in a poor solvent, theory tells us that ensemble-averaged internal distances should plateau to a constant value corresponding to chain length and the density of the collapsed species.[18] The scaling of internal distances for Q_{20} in water and in the globular reference state is found to be consistent with this expectation. The plateau values achieved are in agreement with each other within error. Local length scales, also known as “blob” lengths

are a characteristic of linear flexible polymers.[14] Over this length scale, the scaling of internal distances as a function of sequence spacing is determined primarily by steric interactions, and it is not possible to distinguish good from poor solvents based on conformational equilibria over the "blob" length. Blob lengths can be deduced from the rising part of the curves shown in Figure 2.4 and are found to be about seven to eight residues; consistent with previous findings.[28,29]

Up-and-down topologies in water

Ensemble-averaged angular correlation functions, c_{ij}, provide a way to quantify average topologies adopted by chains in different milieus. This function, analogous to a function proposed by Socci *et al.*,[41] and computed as a function of sequence spacing, is defined as:

$$c_{ij} = \left\langle \left| \cos \Theta_{ij} \right| \right\rangle = \left\langle \left| \frac{\mathbf{l}_i \cdot \mathbf{l}_j}{l^2} \right| \right\rangle \qquad (2\text{-}7)$$

Here, $\mathbf{l}_{i(j)}$ denotes the vector from the backbone nitrogen of residue $i(j)$ to the carbonyl carbon on the same residue, and l is its length. Therefore, Θ_{ij} is the effective angle between the direction of the chain at residues i and j. For chains in a good solvent c_{ij} will decay exponentially as a function of sequence separation $|i\text{-}j|$. Conversely, chains in a poor solvent are under a packing constraint, and on average, the chain will reverse direction. This results in negative values for c_{ij}.

Figure 2.5 shows precisely this behavior. In the EV limit, correlations slowly decay to zero as expected for a worm-like chain. In contrast, the data for

the peptide in water and for the globular reference state are characterized by significant anti-correlation at about five to ten residues of sequence separation. This is the aforementioned mid-range length scale, over which the chain on average turns on itself. Beyond this length scale, correlations decay to zero. The large error bars for the data in water seen in Figure 2.5 are due to two effects: i) every trajectory results in a distinct topology for the globule, and ii) on the timescale of the simulations there is no interconversion between these distinct topologies indicating quenched disorder (see below).

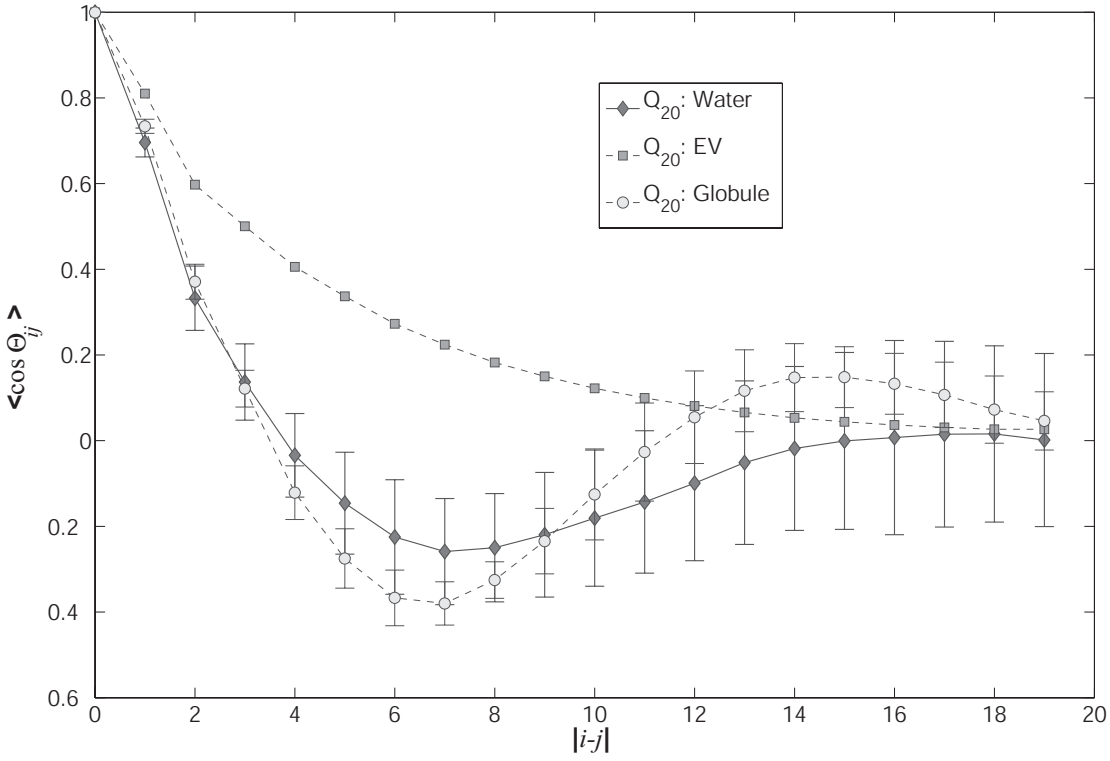

Figure 2.5: **The angular correlation function** (see Equation 2-7) **as a function of sequence separation**. The polypeptide caps are excluded from this analysis. For details on errors see caption to Figure 2.4.

Radial density profiles

Density profiles are another way to characterize the average shape of macromolecules, and form the basis for Lifshitz-type theories for the coil-to-globule transition:[11,17]

$$\rho(r+\Delta r)=\left\langle \sum_{i=1}^{Z_m} \frac{m_i \cdot \left[H\left(r_i - r\right) - H\left(r_i - \left(r+\Delta r\right)\right)\right]}{V\left(r+\Delta r\right) - V(r)} \right\rangle \qquad (2\text{-}8)$$

Here, r_i is the distance of atom i from the molecule's center of mass, m_i is the mass of atom i, Z_m is the number of atoms in the molecule, $V(r)$ is the volume of a sphere with radius r and H is the Heaviside step function.

Figure 2.6 shows that $\rho(r)$ reaches a plateau value for short distances in both the globular reference state and for the peptide in water. The limiting density is ca. 1.2g/cm^3. The most significant difference is in the long distance regime of the density profile. This implies that the peptides in water undergo larger-scale conformational fluctuations than in the globular reference state. The observed plateau value for the density of globules in water and in the LJ reference state is less than that of small, folded proteins.[42] We attribute this difference to the presence of pronounced conformational fluctuations for an IDP such as Q_{20} when compared to stable, folded polypeptides. As may be expected, this discrepancy disappears for longer chain lengths as was established in work comparable to that presented in Chapters IV and V (data not shown). Even for Q_{30}, the limiting density approaches that of small proteins. Conversely, in the EV limit, the density profile is shallow, and reaches a plateau value of about 0.4-0.5g/cm^3. Such a low

value is possible, since chains in the EV limit are characterized by interior cavities of all sizes,[28] and the density is averaged over both void spaces and the chain itself.

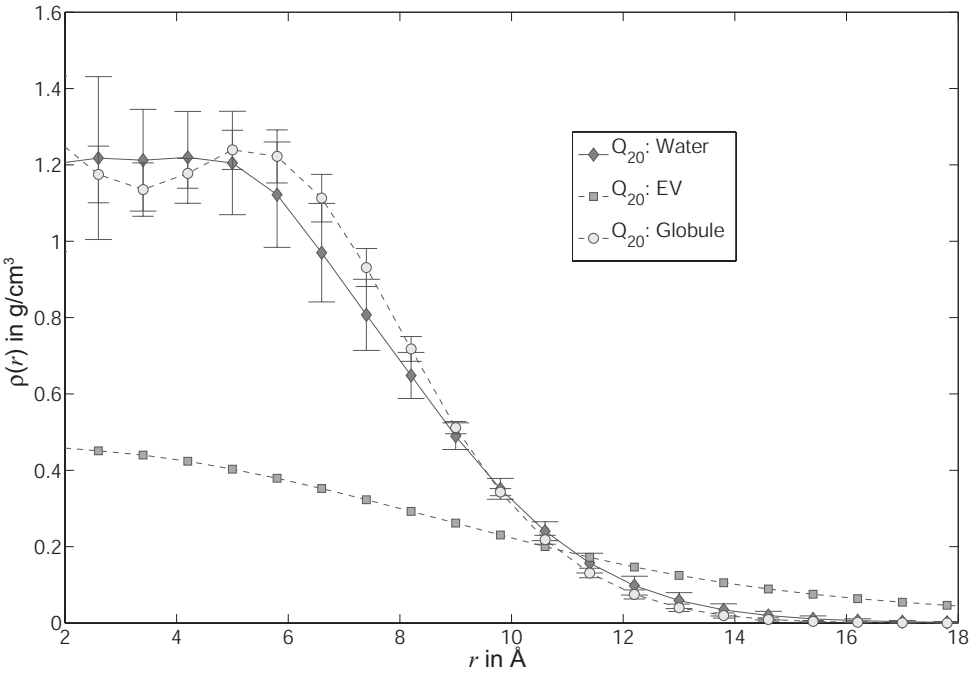

Figure 2.6: **The average density as a function of distance to the center of mass** (see Equation 2-8). For details on errors see caption to Figure 2.4.

Kratky profiles

Finally, Kratky or scattering profiles, $K(q)$,[43] provide a direct connection to experimental data as they are available from small angle X-ray scattering (SAXS) measurements. If we assume homogeneous scattering cross-sections across the molecule, the Kratky profile becomes an effective measure of the peptide's density as a function of a specific length scale:

$$K(q) = Nq^2 \langle P(q) \rangle$$
$$P(q) = \frac{2}{Z_m(Z_m - 1)} \sum_{i=1}^{Z_m} \sum_{j=i+1}^{Z_m} \frac{\sin(qr_{ij})}{qr_{ij}} \qquad (2\text{-}9)$$

Here, the r_{ij} are pairwise atomic distances, Z_m is the number of atoms in the molecule, N is chain length, and q are wavenumbers in units of Å^{-1}. Large peaks in the low and intermediate q-regime ($0.1 \leq q \leq 0.4$) are indicative of compact geometries as they result from a dense collection of scatterers. Conversely, if the Kratky profile is essentially flat with generally low amplitudes, we infer that the scatterers form a loosely packed object with low average density. This is the expected signature for chains in the EV limit.

Figure 2.7 shows that our expectation is again met by the actual data:

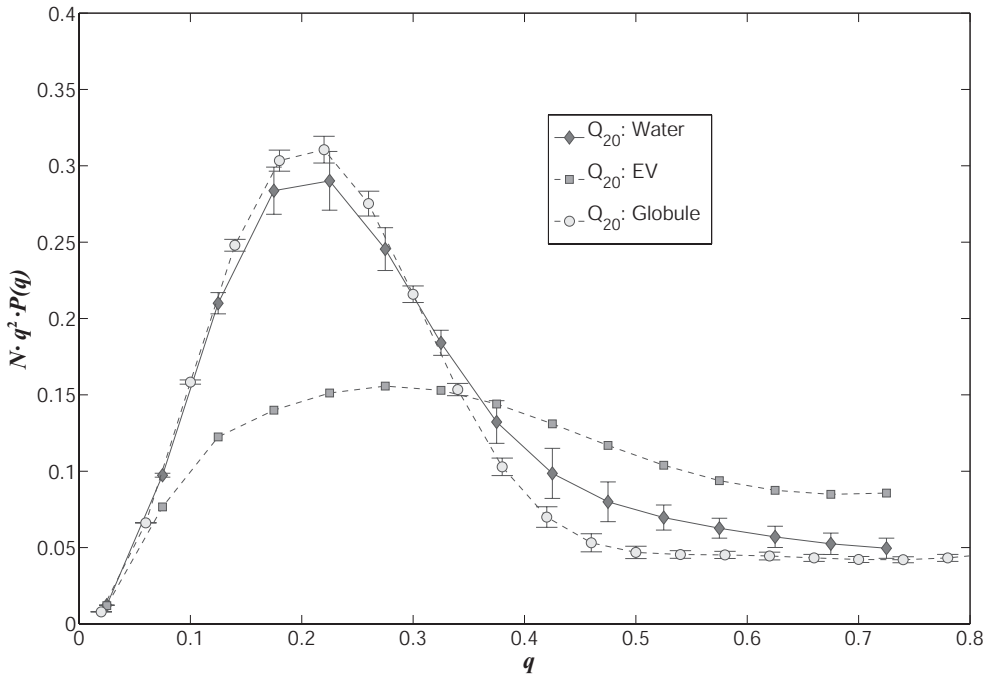

Figure 2.7: **Ensemble averaged Kratky profiles** (see Equation 2-9) **calculated for the three different models**. For details on errors see caption to Figure 2.4.

The profile for the chain in water is very similar to that in the globular reference state, and is undoubtedly distinct from the profile for the EV chain. It is interesting to note that the Kratky profile shows significant quantitative differences between the globular reference and the water data in the high-q regime. This probes differences in local structural propensities between the two ensembles which may ostensibly be accessible to wide angle X-ray scattering experiments.

Moreover, a double-logarithmic plot of $P(q)$ vs. q in principle allows for the direct determination of the scaling exponent for fractal objects given the data for a single chain (much like the internal scaling analysis in Figure 2.4 does in theory). Here, a specific linear regime is fit and the slope is identical to the negative inverse of the scaling exponent. A benefit of this analysis is that Porod's law[43] holds for non-fractal, compact spherical objects of homogeneous density, in which case the slope is expected to be -4. The major drawback is that for small systems the identification of the linear regime is non-trivial. We did, however identify a clear, linear regime for the data in water and obtained a slope of -4.03 (plot not shown), a result consistent with all the data presented in Figures 2.4 through 2.7.

Based on the preceding discussion, we conclude that polymer theory provides us with at least four distinct measures that allow us to establish that water is a poor solvent for Q_{20}. The four quantities we have used to make conclusive analyses are the scaling of internal distances, angular correlation functions to measure average topologies, radial density profiles, and Kratky profiles (closely related to radial density profiles). When these quantities are

computed for data obtained from simulations in explicit water and compared to analysis of simulation data from reference states, we are able to make an unequivocal adjudication regarding the balance between chain-chain and chain-solvent interactions, *i.e.*, solvent quality.

II.4.3. Driving Forces for the Collapse of Polar Polyglutamine in Water

Polyglutamine is a polyamide built of a repeat of secondary amides in the backbone and primary amides in the sidechain. Figure 2.8 shows a comparison of site-site pair correlation functions, $g(r)$, for Q_{20} in water and for aqueous mixtures of dissociated primary and secondary amides:

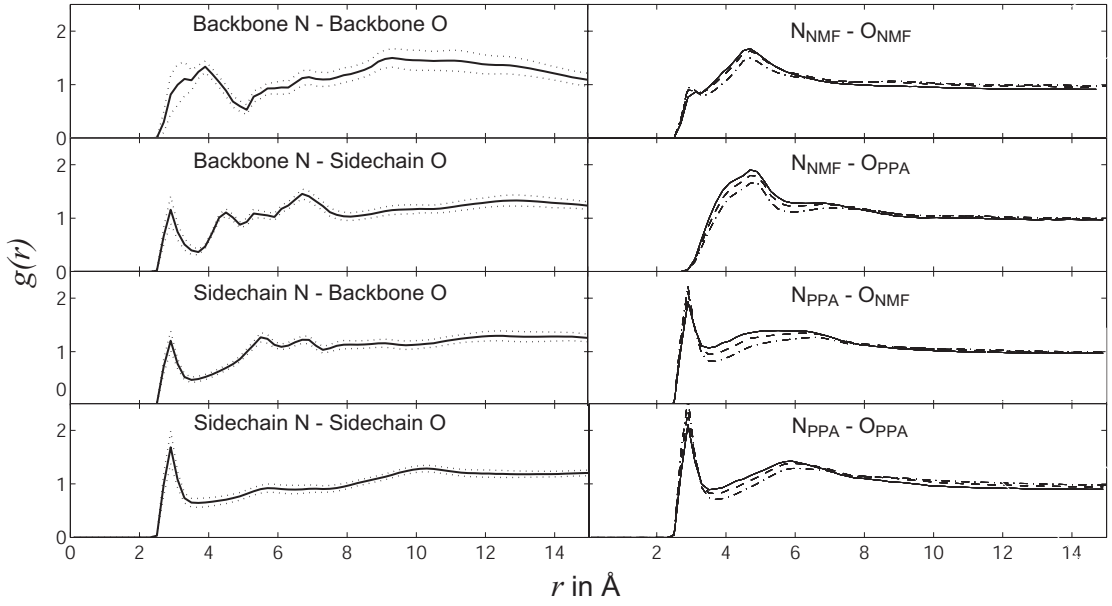

Figure 2.8: **Comparative analysis of pair correlation functions**. The left column shows site-site correlation functions for different atom pairs for Q_{20} in water. The data are normalized by an ideal chain prior (see II.3.6). Dotted lines indicate standard error intervals. The right column shows analogous site-site correlation functions for the

solutions of NMF and PPA in water normalized by an ideal gas prior. Data for three different concentrations are shown (1 m solid curves, 2 m dashed curves, and 4 m dash-dot curves). The sensitivity of the results to amide concentration is small. Standard errors are negligible for these simulations.

We normalized the intra-chain and intermolecular pair correlation functions using different default models because the former is a polymer and the latter is a mixture of freely diffusing molecules. We used an ideal chain model for the polymer and an ideal gas prior for the model compounds. Details are discussed in II.3.6. The model compounds chosen to represent the "dissociated" peptide are *trans*-N-methylformamide (NMF) and propionamide (PPA) mixing freely in solution. NMF, a secondary amide, is an analog of the backbone peptide unit, whereas PPA, a primary amide, is an analog of the polar sidechain of glutamine. We choose NMF instead of *trans*-N-methylacetamide (NMA) due to the incorrect match of total carbon number in the latter.

The first row in Figure 2.8 compares correlation functions between intra-chain backbone donor and acceptor atoms to the site-site correlations between NMF donors (N_{NMF}) and NMF acceptors (O_{NMF}). The first peak around 3Å is pronounced for the polymer and only weakly present for the model compound mixtures in solution. A different scenario holds for the comparison of pair correlations between backbone-donor and sidechain-acceptor atoms to those between N_{NMF} and O_{PPA}, which are shown in the second row of Figure 8. There is a distinct, yet broad peak at 3Å separation in the polymer but general depletion otherwise. For the amide mixtures in solution, the situation is inverted in that

there is relatively strong association at 4-5Å but no short-range peak at ca. 3Å. On the polymer side, the situation is very similar for the inverse pair correlation, *viz.* backbone acceptor and sidechain donor. Again, there is a weak yet distinct peak around 3Å and a general depletion of density for short distances (third row of Figure 2.8). However, for the model compounds, we observe a dominant peak at 3Å followed by a broad second peak in the site-site correlation function for N_{PPA}-O_{NMF}. Finally, there is minimal deviation between pair correlations for the sidechain-sidechain donor-acceptor pair in the polymer and N_{PPA}-O_{PPA} (fourth row of Figure 2.8). For the polypeptide, the correlation function is much smoother than that for other pairs. This is because the sidechains have the most flexibility to rearrange with respect to one another. For both the polymer and the free amides we observe a distinct peak at 3Å.

In summary, we can establish the following changes in the self-association behavior for amides in solution when compared to amides that are part of polyglutamine:

1) For the model compounds in solution, we observe a marked preference for short-range correlations (ca. 3Å) between donor atoms of primary amides (N_{PPA}) and acceptor atoms of secondary amides (O_{NMF}). Interrogation of the inverse pair correlations between sites N_{NMF} and O_{PPA} suggests favorable, solvent-separated intermolecular associations. These differences in donor-acceptor pair correlations are not preserved in the polymer. Instead, both types of pair correlations, *viz.* sidechain donor to backbone acceptor and backbone donor to sidechain acceptor, are equivalent.

2) For the polymer, we observed a general trend that correlation function values are larger than unity for short (ca. 3Å) and long distances (>6Å) but are diminished over medium ranges (3.3Å-6Å). This is due to excluded volume effects, which are absent in the ideal chain model used to normalize the pair correlations (see II.3.6).

3) The two pronounced terms in the polymer are the backbone-backbone and sidechain-sidechain correlation functions, which measure effective interactions between donor and acceptor atoms. Of these two correlation functions, only the pair correlations between backbone units are enhanced vis-à-vis the model compound counterparts. It appears that concatenated backbone units can solvate each other more favorably when compared to free secondary amides. Therefore, our preliminary conclusion is that the main driving force for collapse of polyglutamine in water derives from favorable intra-backbone correlations. This finding appears to be consistent with recent experimental data.[44] There could be multiple sources for enhanced pair correlations. These include hydrogen bonding, the entropic benefits of releasing water molecules into the bulk, and the associated increase in chain packing density.

In the interest of clarity, we reiterate that the intra-polymer and model compound site-site pair correlations were normalized using different default models. Details of the normalization were presented in II.3.6. For the polymer, we used an ideal chain model. This is different from the ideal gas model used as the default model for analyzing distance histograms for model compounds. Therefore, an intra-polymer site-site correlation is meaningful only if the peak or trough in

the pair correlation function is greater than or less than unity, *i.e.*, all enhancements and depletions in intra-polymer pair correlation functions arise due to specific multi-body interactions which either can be repulsive or attractive. The signals should not be misinterpreted as being a consequence of elimination of entropic barriers via chain connectivity.

An alternative approach for making assessments regarding driving forces for collapse is to quantify the contributions of enthalpy and entropy to the free energy change associated with coil-to-globule transitions for polyglutamine. If this transition were to resemble hydrophobic collapse, the driving force would be primarily entropic in nature.[45-49] The data necessary to make judgments regarding entropy and enthalpy are not available from simulations carried out for a single set of solution conditions. Free energy calculations on the solvation of collapsed versus extended states of Q_{20} would be able to address the above issue but are intractable at this point.

II.4.4. Conformational Relaxation Dynamics – Evidence for Glassy Kinetics and Ruggedness of the Energy Landscape

Figure 2.9 shows a checkerboard map of the average root mean square deviation ($\langle \mathrm{RMSD} \rangle_{ij}$) calculated by superposition of all the structures in trajectory j onto the final structure in trajectory i. We find that the diagonal has a significantly lower average RMSD when compared to the off-diagonal elements, *i.e.*, $\langle \mathrm{RMSD} \rangle_{ii} < \langle \mathrm{RMSD} \rangle_{ij}$. This is indicative of two features: first, there is strong residual correlation within each trajectory; second, no pair of trajectories yields

similar final structures. The latter observation clearly establishes the disordered nature of the ensemble.

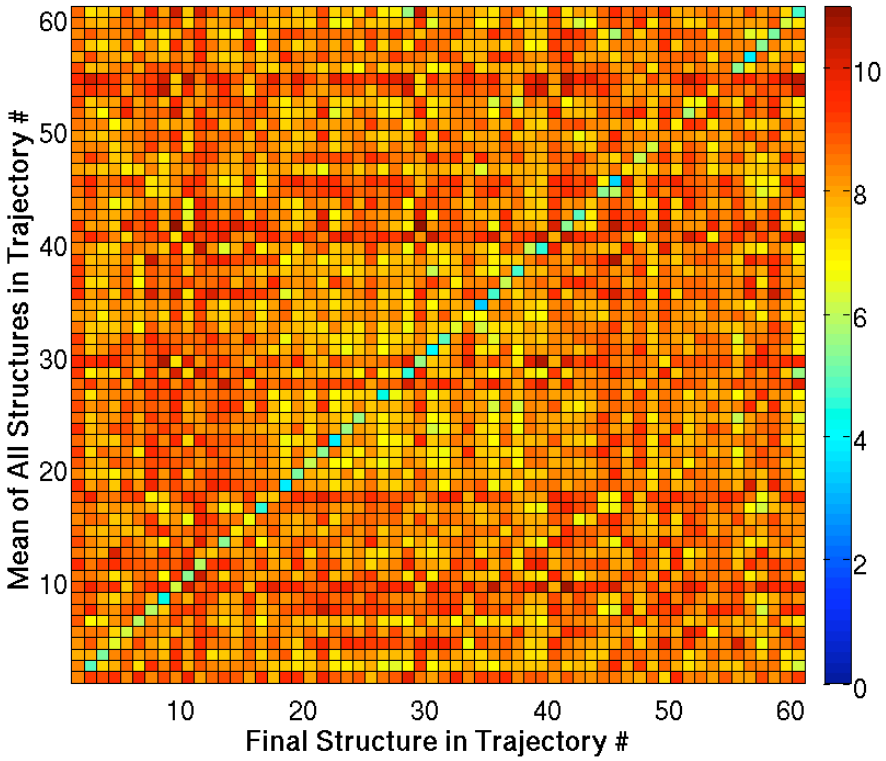

Figure 2.9: **Checkerboard map of the average all-atom RMSD in Å of the structures observed in trajectory *j* (y-axis) from the final structure of trajectory *i* (x-axis)**. This map is by construction not symmetric.

One might argue that inaccurate molecular mechanics force fields as well as the sluggishness of conformational sampling are the primary sources for our observation that the ensemble for polyglutamine is disordered. In other words, the MRMD simulation methodology applied to *any* polypeptide sequence with initial conformations drawn from the EV limit ensemble may yield a similar result. While this skepticism is reasonable, it is also noteworthy that the ensemble dynamics methods of Pande and coworkers, which are similar in spirit to MRMD,

have been used to successfully fold several small two-state proteins and obtain accurate estimates of their folding rates.[50] Therefore, we propose that the congruence between our results and those based on spectroscopic experiments are robust because the homopolymeric nature of polyglutamine provides a reasonable physical basis for its intrinsic disorder. Of course, the concern expressed above can be addressed fully only through application of the MRMD approach to a wide range of sequences that have stable folds as well as to sequences that are predicted to be intrinsically disordered. These types of simulations are computationally challenging and may become feasible with appropriate methodological advances.

In Figure 2.10, we show a comparative analysis of the differences between the time scales for collapse versus the time scales associated with conformational relaxation:

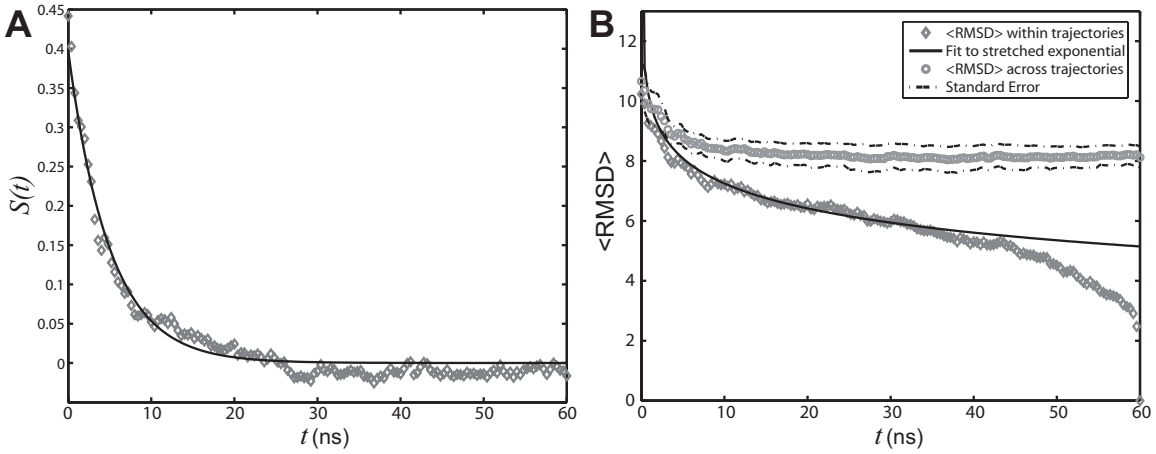

Figure 2.10: Analysis of glassy relaxation dynamics for Q_{20}. Panel A – The time evolution of $S(t)$, a normalized measure of $\langle R_g \rangle$ as a function of time, *t*. The plot also

shows the fit to a single exponential function $S(t) = S_o \exp\left[- t/\tau \right]$ with S_o=0.40 and τ=5ns. The norm of the residuals between the raw data and the exponential function is 0.01. Panel B: RMSD of the structures within a trajectory from their final structure (gray diamonds) is compared to that of the structures within a trajectory to the final structure of other trajectories (gray circles). Standard errors for the former could not be obtained, as there is only one value per trajectory and per time-point. For the cross-term, the 59 values per trajectory and per time-point were pre-averaged and standard errors could be obtained as usual. Data for the average conformational relaxation within a trajectory (gray diamonds) are fit to a stretched exponential function of the form described in the text. This is shown as the solid curve in the plot. Deviations from the stretched exponential function are largest for the earliest time points, t<5ns and for the last 10ns interval. The former is explained by the rapid collapse over short time scales, while the latter is entirely due to our choice of the final snapshot of the trajectory as the reference snapshot for analyzing conformational relaxation.

In Panel A, we plot $S(t) = \frac{\left[\langle R_g \rangle (t) - \langle R_g \rangle \right]}{\langle R_g \rangle}$ as a function of time. A single exponential fit for the decay of $S(t)$ versus t is also shown. This function, $S(t) = S_o \exp\left[-\frac{t}{\tau} \right]$ has the parameters S_o=0.40 and τ=5ns. In each of the trajectories, collapse from the relatively extended starting conformations, which are extracted from the EV ensembles, is found to be a rapid process and occurs within a time scale of approximately 5ns, which is shorter than the equilibration times (11ns) used in our analysis of MRMD data. This observation is robust

across all trajectories. In the interest of clarity, we have added data from the equilibration periods. This was done for the analysis reported in Figure 2.10 alone. For all other figures, only data from the production runs were used.

Although collapse is rapid, conformational relaxation is considerably slower. Panel B of Figure 2.10 shows the time evolution of the average RMSD for superposition of structures within a trajectory i to the final structure of trajectory i, *i.e.*, $\langle \mathrm{RMSD}(t) \rangle_{\mathrm{self}}$. The temporal evolution of this parameter is described using a stretched exponential function: $\langle \mathrm{RMSD}(t) \rangle_{\mathrm{self}} = R_o \exp[-(t/\tau)^{\beta}]$, with R_o=22Å and β=0.15; Here, τ is set to be 5ns – the time scale for collapse. The stretched exponential function, also known as the Kohlrausch-Williams-Watts (KWW) function (with $0 \leq \beta \leq 1$), is used to describe structural relaxation in glassy systems (below the glass transition temperature).[27,51-53] If β assumes small values, then the system has access to a broad and heterogeneous distribution of relaxation times.[51] Our discovery that conformational relaxation of Q_{20} follows non-exponential kinetics with a fairly small value of β is consistent with the postulate that distinct collapsed structures are likely to be of equivalent stability on account of the homopolymeric nature of polyglutamine, *i.e.*, the energy landscape is rugged for Q_{20} in water at T=298K and P=1bar.

There are two predicted features for rugged energy landscapes: The first is slow, non-exponential relaxation within distinct basins, which is best described using a KWW function.[27,51,52] Secondly, there should be evidence of even slower interconversion between distinct basins.[51] Evidence for the latter is also shown in

Panel B of Figure 2.10. Here, we track the temporal evolution of $\langle \mathrm{RMSD}(t) \rangle_{\mathrm{cross}}$, which is the average RMSD for superposition of a snapshot from trajectory i upon the final structure of trajectory j, where $j \neq i$. The desired average is calculated over all unique pairs of trajectories (i) and final structures (j). We find that, once the chain is collapsed, $\langle \mathrm{RMSD}(t) \rangle_{\mathrm{cross}}$ shows no significant time dependence over the remaining time scale of 50ns. The time dependencies of both $\langle \mathrm{RMSD}(t) \rangle_{\mathrm{cross}}$ and $\langle \mathrm{RMSD}(t) \rangle_{\mathrm{self}}$ taken together are interpreted as follows: Although collapse is rapid and the $\langle R_{\mathrm{g}} \rangle$-values across trajectories are similar to each other, each trajectory samples a distinct family of globular conformations, and there is no obvious interconversion between the distinct globules over the 50ns time scale.

Our MRMD approach provides reliable information regarding global, polymeric order parameters because this information is converged and roughly equivalent across all trajectories. Conversely, any analysis of specific structural propensities would yield mostly unreliable information because this requires interconversion between distinct conformational basins. To achieve this, each independent trajectory in the MRMD approach will need to be extended into the μs-range or longer. Perhaps, an increase in the number of independent trajectories will be necessary as well. The impact of conformational heterogeneity and diminished conformational averaging is seen in the large error bars for the angular correlation function (Figure 2.5). This measure probes local

conformational propensities as well as global properties and is therefore most sensitive to the quality of statistics we gather.

II.5. Summary and Conclusions

We have analyzed MRMD simulations for a single polypeptide chain, Q_{20}, in water. Our analysis – combined with polymer physics theories – and comparison to data from reference simulations allows us to conclude that Q_{20} in water has all the characteristics of a chain in a poor solvent (Figures 2.2-2.7). The physics of homopolymers allows us to generalize and conclude that water is a poor solvent for polyglutamine, *i.e.*, at infinite dilution these systems form disordered globules and at finite concentrations the stable thermodynamic state will be the phase-separated aggregate.[14,54] Implications of the poor solvent nature of aqueous solvents for the mechanism of aggregation have been discussed in detail[19] and will be elaborated upon later (see Chapter IV in particular).

Polymer theory helps in making robust predictions

We borrowed the methods for analyzing conformational equilibria from the polymer physics literature.[11,14,16,18,55] The motivation was to ask if the analysis of simulation data for a single chain length could lead to robust assertions about solvent quality. We showed that this is possible using comparative analysis of specific "order parameters".[17] Of particular relevance is the scaling of internal distances because it obeys a rigorous scaling law for fractal objects, *i.e.*, chains in good and θ-solvents. Departure from a scaling law for this quantity must mean

that the solvent is poor. Finite size effects limit the usefulness of such a measure only if the chain length drops below the "blob" length of seven to eight residues since in this regime local structure overrides the mean polymeric behavior.[14] The presence of two distinct length scales, *viz.* the blob length and a generic length, also means that the conclusions obtained from our analysis for N=20 are robust and valid for all chain lengths N>20. This point is emphasized in the development of modern theories for homopolymers[11,14,17] and in the previous work by Crick *et al.*[19] who showed that the poor solvent scaling of chain size with length is obeyed for all lengths $N \geq$15. Our analysis was feasible due to low-sequence complexity, *i.e.*, the homopolymeric nature of polyglutamine and the appropriate choice of chain length (longer than the blob length). The analysis methods are likely to be of general relevance for quantitative characterization of conformational equilibria for IDPs because many of these sequences are deficient in hydrophobic residues and are of sufficiently low sequence complexity.[3,8]

Why is water a poor solvent for glutamine-rich peptides?

Combining experimental studies and our computational results, there remains little doubt that water is in fact a poor solvent for glutamine-rich peptides. These peptides are assumed to be in a "random-coil" state, the implication being that the ensemble is consistent with that of highly denatured proteins. Our results suggest that the absence of a consensus experimental signal in CD experiments[20,38,56] is the result of a different type of disorder, *i.e.*, of a heterogeneous ensemble of globular conformations. Given the polar nature of

the sidechain, and the infinite solubility of small amides in water, it is obvious that the solvation behavior changes upon transitioning from amides in water to a polyamide in water. To be able to compare the two cases, we remove effective concentration as an obvious factor by appropriate normalization (Figure 2.8). We conclude that the short-range steric and topological constraints in the polymer alter the solvation behavior primarily for the backbone unit, *i.e.*, the secondary amides are more favorably solvated by themselves than by water. As a result, the chain collapses and minimizes its interface with water.

However, the above does not imply that these peptides behave like classical hydrophobic solutes, such as polyethylene. At this point, we are unable to adjudicate the nature of the collapse transition since we only have simulations of conformational equilibria for a single set of solution conditions. However, the qualitative result in and of itself appears to be robust. More recent work has shown that another archetypical IDP free of any hydrophobic residues, *viz.* polyglycine, similarly collapses in water.[57] Given the fact that polyglycine is also a polyamide but has no sidechains, the hypothesis that the backbone amides appear to be the driving force behind this somewhat counterintuitive phenomenon of “polar collapse” appears reasonable. Further work in this direction is currently being carried out in an attempt to quantify both the physicochemical origin and the structural signatures of polar collapse (Wyczalkowski and Pappu, unpublished).

Implications for the feasibility of future computational studies on polyglutamine aggregation using the MRMD protocol presented here

Ultimately, our interest lies in characterizing the process of aggregation of peptides of the type studied here. As is clear from Figures 2.3 through 2.7, the standard errors for most of the data concerning the conformational equilibrium of the peptide are relatively large considering the investment of computational resources. This is a direct consequence of the very long interconversion times for different globular states of these peptides as detailed in Figures 2.9 and 2.10. It might be possible to reduce the computational cost for this particular system via enhanced sampling techniques.[58-60] It was demonstrated that a global coordinate such as R_g can be used as a *bona fide* reaction coordinate in an umbrella sampling approach.[57] However, the fundamental problem of increasing system sizes remains unsolved. With an explicit representation of the solvent, the number of solvent molecules will increase linearly with volume. The required box lengths for studies of monomeric polyglutamine will ideally increase linearly with chain length. Ultimately, this yields a cost dependency on chain length of N^3 even if we optimistically assume linear scaling of CPU time with number of atoms in the system.

Given current computing resources and given the manifold of sequences researchers will ultimately be interested in simulating, mean-field representations of the solvent are therefore a necessity for computer simulations to address questions pertaining to the conformational equilibria and aggregation of disease-associated IDPs such as polyglutamine. A development of a novel mean-field

representation of water is presented in Chapter III and Chapters IV through VI detail its application to the polyglutamine aggregation problem. However, additional help with sampling is needed. The popular replica exchange (REX) method[61,62] could theoretically have been used here. It employs high temperature replicas to enhance conformational re-arrangement via increased rates of barrier-crossing. While this is useful in theory, we would suffer from the fact that we would need multiple replicas for each temperature. This is unavoidable for disordered systems such as polyglutamine in water; therefore the required resources would actually *increase*. A much more straightforward application of the REX technique emerges if we are interested in quantities as a function of the control parameter, typically temperature, *and* if the bulk properties of the solvent bath do not depend drastically on the control parameter. Neither condition was true for the scope of this chapter. Both *are* true for Chapters IV, V, and VI which routinely employ the REX method.

As is the case in most molecular simulations of biomolecules, the choice of the force-field will determine the details of simulation results.[63] Since all force-fields share similar features, our analysis methods applied to simulation data gathered using different force-fields will in all likelihood lead to the conclusion that water is a poor solvent for polyglutamine. However, details such as the length scale for collapse transitions and the stability of the collapsed states might vary from one force-field to the next. The OPLS-AA/L force field[30] employed in this study has not been parameterized with this application in mind. It is worth pointing out that its parameterization paradigm rests mostly on small molecule

data; hence, better transferability and accuracy for basic physicochemical properties can be expected than for other common force fields. Nevertheless, a comparative study across multiple force fields appears desirable. While such studies have become more common in recent years,[64-68] they are still prohibitively expensive. For the data presented here, we used ca. 1200 CPU days on a single 2.6Ghz Intel Conroe Core with the fastest, freely available simulation engine, *viz.* GROMACS. Maybe more than anything else, this raw number clearly points out the computational cul-de-sac the MRMD approach presents for studying a system of this complexity. Following the preceding discussion, this is especially true given the need to repeatedly demonstrated *both* accuracy *and* reliability.

II.6 Bibliography

1. Vitalis, A.; Wang, X.; Pappu, R. V. Biophys J 2007, 93(6), 1923-1937.

2. Wang, X. L.; Vitalis, A.; Wyczalkowski, M. A.; Pappu, R. V. Proteins: Struct Funct Bioinf 2006, 63(2), 297-311.

3. Dunker, A. K.; Brown, C. J.; Obradovic, Z. Adv Protein Chem 2002, 62, 25-49.

4. Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradovic, Z. Biochemistry 2002, 41, 6573-6582.

5. Uversky, V. N. Protein Sci 2002, 11, 739-756.

6. Dyson, H. J.; Wright, P. E. Nature Rev Mol Cell Biol 2005, 6, 197-208.

7. Weathers, E. A.; Paulaitis, M. E.; Woolf, T. B.; Hoh, J. H. FEBS Lett 2004, 576(3), 348-352.

8. Weathers, E. A.; Paulaitis, M. E.; Woolf, T. B.; Hoh, J. H. Proteins: Struct Funct Bioinf 2007, 66(1), 16-28.

9. Derkatch, I. L.; Uptain, S. M.; Outeiro, T. F.; Krishnan, R.; Lindquist, S. L.; Liebman, S. W. Proc Natl Acad Sci U S A 2004, 101(35), 12934-12939.

10. Bright, J. N.; Woolf, T. B.; Hoh, J. H. Prog Biophys Mol Biol 2001, 76, 131-173.

11. Grosberg, A. Y.; Khokhlov, A. R. Statistical Physics of Macromolecules; AIP Press: New York, 1994.

12. Flory, P. J. Principles of Polymer Chemistry; Cornell University Press: Ithaca and London, 1953.

13. Chan, H. S.; Dill, K. A. Annu Rev Biophys Biophys Chem 1991, 20, 447-490.

14. Rubinstein, M.; Colby, R. H. Polymer Physics; Oxford University Press: Oxford and New York, 2003.

15. Reddy, G.; Yethiraj, A. Macromolecules 2006, 39(24), 8536-8542.

16. Steinhauser, M. O. J Chem Phys 2005, 122(9).

17. Grosberg, A. Y.; Kuznetsov, D. V. Macromolecules 1992, 25(7), 1970-1979.

18. Imbert, J. B.; Lesne, A.; Victor, J. M. Phys Rev E 1997, 56(5), 5630-5647.

19. Crick, S. L.; Jayaraman, M.; Frieden, C.; Wetzel, R.; Pappu, R. V. Proc Natl Acad Sci U S A 2006, 103(45), 16764-16769.

20. Chen, S.; Berthelier, V.; Yang, W.; Wetzel, R. J Mol Biol 2001, 311(1), 173-182.

21. Scherzinger, E.; Lurz, R.; Turmaine, M.; Mangiarini, L.; Hollenbach, B.; Hasenbank, R.; Bates, G. P.; Davies, S. W.; Lehrach, H.; Wanker, E. E. Cell 1997, 90(3), 549-558.

22. Krull, L. H.; Wall, J. S. Biochemistry 1966, 5(5), 1521-1527.

23. Wolfenden, R. Biochemistry 1978, 17(1), 201-204.

24. Wolfenden, R.; Andersson, L.; Cullis, P. M.; Southgate, C. C. B. Biochemistry 1981, 20(4), 849-855.

25. Chen, S. M.; Ferrone, F. A.; Wetzel, R. Proc Natl Acad Sci U S A 2002, 99(18), 11884-11889.

26. Chuang, J.; Grosberg, A. Y.; Tanaka, T. J Chem Phys 2000, 112(14), 6434-6442.

27. Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Proteins: Struct Funct Genet 1995, 21(3), 167-195.

28. Tran, H. T.; Pappu, R. V. Biophys J 2006, 91(5), 1868-1886.

29. Tran, H. T.; Wang, X. L.; Pappu, R. V. Biochemistry 2005, 44(34), 11369-11380.

30. Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. J Phys Chem B 2001, 105(28), 6474-6487.

31. Lindahl, E.; Hess, B.; van der Spoel, D. J Mol Model 2001, 7(8), 306-317.

32. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. J Chem Phys 1983, 79(2), 926-935.

33. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. J Chem Phys 1984, 81(8), 3684-3690.

34. Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. J Comput Chem 1997, 18(12), 1463-1472.

35. Miyamoto, S.; Kollman, P. A. J Comput Chem 1992, 13(8), 952-962.

36. Onsager, L. J Am Chem Soc 1936, 58(8), 1486-1493.

37. Dima, R. I.; Thirumalai, D. J Phys Chem B 2004, 108(21), 6564-6570.

38. Masino, L.; Kelly, G.; Leonard, K.; Trottier, Y.; Pastore, A. FEBS Letters 2002, 513(2-3), 267-272.

39. Schäfer, L. Excluded Volume Effects in Polymer Solutions as Explained by the Renormalization Group; Springer: Berlin, 1999.

40. Ding, F.; Jha, R. K.; Dokholyan, N. V. Structure 2005, 13(7), 1047-1054.

41. Socci, N. D.; Bialek, W. S.; Onuchic, J. N. Phys Rev E 1994, 49(4), 3440-3443.

42. Fischer H., P., I., Craievich, A.F. Protein Sci 2004, 13, 2825-2828.

43. Glatter, O.; Kratky, O. Small Angle X-Ray Scattering; Academic Press: London, 1982.

44. Moglich, A.; Joder, K.; Kiefhaber, T. Proc Natl Acad Sci U S A 2006, 103(33), 12394-12399.

45. Chandler, D. Nature 2005, 437(7059), 640-647.

46. Southall, N. T.; Dill, K. A.; Haymet, A. D. J. J Phys Chem B 2002, 106(3), 521-533.

47. Paulaitis, M. E.; Garde, S.; Ashbaugh, H. S. Curr Opin Coll Interf Sci 1996, 1(3), 376-383.

48. Athawale, M. V.; Goel, G.; Ghosh, T.; Truskett, T. M.; Garde, S. Proc Natl Acad Sci U S A 2007, 104(3), 733-738.

49. Schmid, R. Monatsh Chem 2001, 132(11), 1295-1326.

50. Pande, V. S.; Baker, I.; Chapman, J.; Elmer, S. P.; Khaliq, S.; Larson, S. M.; Rhee, Y. M.; Shirts, M. R.; Snow, C. D.; Sorin, E. J.; Zagrovic, B. Biopolymers 2003, 68(1), 91-109.

51. Sastry, S.; Debenedetti, P. G.; Stillinger, F. H. Nature 1998, 393(6685), 554-557.

52. Phillips, J. C. Rep Prog Phys 1996, 59(9), 1133-1207.

53. Thirumalai, D.; Mountain, R. D. Phys Rev E 1993, 47(1), 479-489.

54. Raos, G.; Allegra, G. J Chem Phys 1997, 107(16), 6479-6490.

55. Grosberg, A. Y.; Kuznetsov, D. V. J Phys II 1992, 2(6), 1327-1339.

56. Bennett, M. J.; Huey-Tubman, K. E.; Herr, A. B.; West, A. P.; Ross, S. A.; Bjorkman, P. J. Proc Natl Acad Sci U S A 2002, 99(18), 11634-11639.

57. Tran, H. T.; Mao, A.; Pappu, R. V. J Am Chem Soc 2008, 130(23), 7380-7392.

58. Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. J Comput Chem 1995, 16(11), 1339-1350.

59. Sheinerman, F. B.; Brooks, C. L. J Mol Biol 1998, 278(2), 439-456.

60. Roux, B. Comput Phys Commun 1995, 91(1-3), 275-282.

61. Sugita, Y.; Okamoto, Y. Chem Phys Lett 1999, 314(1-2), 141-151.

62. Nymeyer, H.; Gnanakaran, S.; Garcia, A. E. Methods Enzymol 2004, 383, 119-149.

63. Sorin, E. J.; Pande, V. S. J Comput Chem 2005, 26(7), 682-690.

64. Hu, H.; Elstner, M.; Hermans, J. Proteins: Struct Funct Genet 2003, 50(3), 451-463.

65. Zagrovic, B.; Pande, V. S. J Am Chem Soc 2006, 128(36), 11742-11743.

66. Gnanakaran, S.; Garcia, A. E. Prot Struct Funct Bioinf 2005, 59(4), 773-782.

67. Zaman, M. H.; Shen, M. Y.; Berry, R. S.; Freed, K. F.; Sosnick, T. R. J Mol Biol 2003, 331(3), 693-711.

68. Khalili, M.; Wales, D. J. J Chem Theory Comput 2009, 5(5), 1380-1392.

CHAPTER III. DEVELOPMENT OF A NOVEL IMPLICIT SOLVENT MODEL TO FACILITATE SIMULATIONS OF THE ASSOCIATION OF DISORDERED POLYPETIDES RICH IN GLUTAMINE

III.1. Preamble

As the results in Chapter II demonstrate, a brute-force computational strategy resting entirely on what appears to be the most adequate and accurate model for the problem of polyglutamine aggregation, *i.e.*, MD in explicit solvent, was and is infeasible. Specifically, the strict coupling to a temporal evolution of the system and the extremely poor scaling behavior with system size present fundamental obstacles even for obtaining converged properties of the relatively short peptide Q_{20}. However, as is outlined in Chapter I, the relevant physiological length scale extends well beyond that range. More importantly, at representative and hence relevant concentrations (referring to typical experimental conditions in both *in vivo* and *in vitro* settings) even the minimal unit of aggregation, namely the dimer, requires simulation systems of sizes completely inaccessible to an all-atom representation of the solvent.

We therefore endeavored to coarse-grain, *i.e.*, to reduce the represented number of degrees of freedom in the system, loosely defined as μM solutions of polypeptides rich in glutamine. Significant time was spent in late 2005 and early 2006 developing a formalism ultimately resting on an inverse Boltzmann strategy.[1] Here, knowledge-based potentials are extracted from simulation databases and defined over coordinates presumed to be statistically independent,

i.e., separable from the viewpoint of the system Hamiltonian. Ultimately, this strategy failed even when considering collective coordinates obtained using statistical decomposition approaches. In particular, we tried to obtain linear combinations of the dihedral angles as collective coordinates. The hypothesis was that via a technique called independent component analysis, ICA,[2] we would arrive at a set of collective coordinates for which the knowledge-based potentials are by definition separable. In such a scenario, application of inverse Boltzmann techniques is trivial. The approach gave unsatisfactory results primarily due to the following fundamental weakness: the theory underlying ICA (and the more popular PCA) is entirely linear, whereas the actual correlations in the data showed substantial non-linearities.

It may actually be called a fortunate event that the knowledge-based approach failed so rigorously. It motivated us to pursue an avenue explored previously (summer of 2005) with renewed emphasis. Fundamentally, an alternative coarse-graining strategy lies in an attempt to conceptualize the physics of the process of interest and to describe that conceptualization in quantitative terms. A model emerging from such an approach might be called semi-empirical: parameter optimization is undoubtedly needed, but all parameters relate to physically intuitive properties. Molecular mechanics force fields are a very good example for a semi-empirical method: for instance, rather than describing Coulombic interactions at the level of electronic structure, they use a point charge model to represent static multipole interactions. The partial charges for individual atoms remain to be fit, but their physical meaning is

straightforward. This has the important advantage that auxiliary calculations and experiments can guide the parameterization. In case of partial charges, those can be quantum mechanics calculations in conjunction with experimentally determined dipole moments.

Implicit solvent models and their underlying theories such as continuum electrostatics theories attempt to follow the semi-empirical paradigm outlined in the previous paragraph. The difference is that here interactions with the solvent, typically aqueous, are attempted to be captured in a mean-field fashion. Since the tradeoff of computational efficiency and accuracy did not appear favorable for many of the established continuum solvation models, we developed a novel one, which – of course – at the same time remains deeply rooted in the work of others. The remainder of this chapter is mostly identical to the work we published on the description and testing of our novel implicit solvent model termed ABSINTH (for self-**A**ssembly of **B**iomolecules **S**tudied by an **I**mplicit, **N**ovel, and **T**unable **H**amiltonian).[3]

III.2. Introduction to Implicit Solvent Models

Generally speaking, computer simulations of biomolecules complement experimental methodologies by providing a detailed representation of the system of interest. They allow for analysis of novel quantities and lead to insights regarding the mechanisms and driving forces underlying experimentally observed phenomena.[4] Common simulation methodology and popular molecular mechanics force fields are usually designed to work with explicit water models,

i.e., all solvent molecules in the system must be represented explicitly in atomic detail. As mentioned in III.1, this can become prohibitively expensive for biological phenomena such as self-assembly or even the unfolding of a single protein molecule. These processes require spontaneous fluctuations that span multiple length and time scales. The idea of representing solvent as a continuum in particular for studying large-scale phenomena has therefore retained appeal within the simulation community.[5] If one uses an implicit / continuum model for solvation, the computational cost of a single energy or force calculation will, in theory, scale with the number and size of the biomolecules of interest rather than with the spatial dimensions of the simulation system.

Part of the motivation for developing a *new* implicit solvation model emerged and continues to emerge from growing interest in the topic of intrinsically disordered proteins (IDPs), amongst them polyglutamine (see Chapter II). IDPs do not fold into well-defined, ordered tertiary structures under physiological conditions. Disorder prevails under non-denaturing conditions and amino acid sequence encodes the propensity to be disordered. Recent data from simulations using explicit solvent models and fluorescence-based experiments show that archetypal polar IDPs such as polyglutamine,[6] the N-domain of the yeast prion protein Sup35, and glycine-serine block copolypeptides[7] form an ensemble of disordered, collapsed structures in water. Disorder in these systems is not a consequence of the inability to collapse; rather it reflects a lack of sequence specificity for a unique collapsed structure, *i.e.*, a fold.

Naturally, the domain of parameterization for almost all common force fields has been that of well-folded structures. It has been argued that strong biases toward canonical polypeptide secondary structure exist in different force fields. An example is the strong α-helical tendency of earlier incarnations of the AMBER force field.[8,9] This implies that an inadvertent bias against disorder is built into most of these force fields. They therefore appear unsuitable to capture properly the subtle interplay between conformational entropy and enthalpy involved in an IDP adopting transient, partially folded structures when interacting with suitable binding partners. Similarly, they appear unsuitable in elucidating the role of canonical secondary structure in driving or modulating the aggregation of an IDP like polyglutamine.

As pointed out above, for such studies we need to carry out large-scale simulations in atomistic detail; hence, we require highly efficient molecular simulations. Furthermore, since the primary objective is to describe conformational ensembles in terms of coarse-grained order parameters, it is reasonable to pursue the development of implicit solvent models that emphasize speed with some tradeoff in fine-grained accuracy. For example, the type of model we have developed here would not be ideal for predicting the three-dimensional structures of proteins to very high accuracy. Instead, it is intended to be useful for identifying the native-state basin in a coarse-grain manner while also providing quantitatively accurate assessments regarding competing conformational basins. The latter is especially useful for understanding how

spontaneous fluctuations lead to disorder-mediated functional interactions as well as deleterious interactions such as protein aggregation.

Prior to summarizing the features of the new model, we first review the features that underlie existing approaches for modeling solvent in an implicit manner. Methods based on the Poisson-Boltzmann (PB)[10] equation are regarded as the most accurate implicit solvent models in terms of electrostatics. The Poisson equation is based on the assumption of a dipolar continuum for the solvent. The polar contribution to the solvation free energy of a biomolecule is modeled as the mean-field response of a dipolar continuum to the formation of a set of point charges within a low dielectric cavity that is in turn embedded in a high-dielectric medium. In the PB equation, the continuum is extended by a mobile, Boltzmann-distributed charge density. With current computing power, both the Poisson and Poisson-Boltzmann equations can be solved numerically even for very large systems to a high level of accuracy.[11] This provides a strategy to estimate the solvation free energy of individual biomolecular conformations or specific, large-scale assemblies. However, such calculations remain prohibitively expensive for most simulation purposes where one needs large numbers of independent evaluations of solvation free energies for the system of interest.[12] Additionally, while the polar contributions to the transfer of a complex solute into the continuum and the dielectric screening of polar solute-solute interactions are modeled accurately, PB methods cannot address the non-polar part of the transfer process. In principle, this is achieved by addition of a non-polar term as described below. In practice, PB methods are not typically used in simulations of

biomolecules; rather they are used for interrogating solvation free energies of static structures.

Generalized Born (GB) models[5] are an analytical approximation to PB models. In the GB surface area (GB/SA) variants, the non-polar contribution to the transfer process is represented by a surface-area based term, while the electrostatic contribution is based on an analytical expression. The Born equation, generalized to account for the macromolecular environment, describes the charging process for individual sites. Cross-terms represent the modulation of polar interactions by the dipolar continuum and by the protein.[13] Most of the deviations of the GB approach from the PB model to modeling electrostatics can be attributed to inaccurate Born radii, which result from approximations to the appropriate integrals.[14] Additional errors arise because reaction-field effects are ignored.[15]

In the earliest incarnations of GB/SA models, the non-polar treatment relied on the solvent-accessible surface area (SASA) to describe cavitation.[16] It is well-known, however, that the validity of the SASA to describe hydrophobic solvation only holds beyond a certain length scale and that the solvent-accessible volume (SAV) provides a better metric for rough surfaces with high curvature.[17-20] Moreover, dispersion terms describing favorable non-polar interactions between solute and solvent have also been shown to be relevant.[20,21] Consequently, significant improvements in the non-polar treatment in GB models have been achieved by adding a volume-dependent dispersive term to the SASA-dependent cavitation term.[22]

It should be noted that both PB and GB methods can suffer from surprisingly poor performance when compared to explicit solvent calculations depending on the system. GB models become ineffective if the calculation of Born radii needs to be repeated frequently as would be the case in Monte Carlo (MC) simulations where large conformational changes can occur rapidly. Conversely, PB methods require numerical solutions of the Poisson-Boltzmann equation and remain comparably slow despite significant advances in the available technology.[10] Often, in both PB and GB models, there can be a tradeoff of accuracy for speed,[23] which might be appropriate for certain systems but not in general. It is also noteworthy that GB models are usually calibrated with respect to PB models and not with respect to calculations in explicit solvent. This leads to internal consistency between the two models. However, weaknesses due to the assumption of a dipolar continuum prevail in both models, and this weakness[24] is emphasized by the hypersensitivity of PB/GB models to the definition of the dielectric boundary.[10,25]

There are other, simpler versions of implicit solvent models. These yield qualitatively correct results and have been used to extend the time scale in molecular dynamics (MD) simulations well into the μs-range. Caflisch and coworkers[26,27] have employed a SA-based term to capture the mean-field interaction of the solute with the solvent and a simple distance-dependent dielectric to describe the modulation of polar interactions by the continuum. The EEF1 model by Lazaridis and Karplus[28] follows a paradigm which differs fundamentally from that of PB/GB(SA) models. Here, the transfer process is

decomposed into a direct mean-field interaction and a screening term rather than into polar and non-polar contributions as is the case in PB/GB(SA) models. The treatment of the direct mean-field interaction (DMFI) is designed to reproduce experimental transfer free energies from vacuum into aqueous solution for small functional groups according to a decomposition scheme proposed by Privalov and Makhatadze.[29] The sum of these contributions determines the maximal, net solvation free energy for the entire biomolecule. This sum is reduced from reference values if the accessibility of the sites is less than maximal, *i.e.*, if other solute atoms shield solvation sites from the continuum. The EEF1 model does not rely on the popular SASA-metric to determine accessibility. Instead, it employs a Gaussian, volume-based term corresponding to the SAV. In its original implementation, EEF1 used a simple distance-dependent dielectric to describe the screening of Coulombic interactions. This was later revised to include an exposure-dependent component.[30]

In designing our implicit solvation model, we aimed to maximize efficiency and accuracy with respect to the target applications while also offering the ability to tune the model and make it more versatile. The result is a model we refer to as ABSINTH, which stands for self-**A**ssembly of **B**iomolecules **S**tudied by an **I**mplicit, **N**ovel, and **T**unable **H**amiltonian. In ABSINTH, the transfer process of a solute into the continuum is written as the sum of two terms, *viz*. a DMFI, and a term used to model the screening of polar interactions. The solute molecule is decomposed into sets of distinct solvation groups. The DMFI is written as a sum of contributions from each of the solvation groups, which are analogs of model

compounds. SAV fractions (η) are used as the metric for solvent accessibility. Electrostatic interactions are treated using charge groups to eliminate spurious, short-range electrostatic interactions. Continuum-mediated screening of these interactions is treated as a purely environmental term with no explicit distance-dependence using a framework similar to the one used for the DMFI. Finally, we do not use torsional potentials and both Lennard-Jones (LJ) parameters as well as partial charges are treated as modular entities, *i.e.*, they are not co-dependent. As discussed below, the model offers parameters that allow one to tune the cooperativity of transitions between fully solvated and fully desolvated states, although we have not fully explored this feature in the present work.

To summarize, in ABSINTH both the polar and non-polar parts of the transfer process are treated simultaneously using reference free energies of solvation for the solvation groups, which is fundamentally different from the approach taken by PB and GB models. Differences between EEF1 and ABSINTH arise in the way we measure the solvent accessibility. We introduce a generalized, stretched sigmoidal function to compute solvation states from solvent accessibilities. We also depart from EEF1 in the choice of solvation groups; we use larger model compounds, thereby using experimental data directly without relying on empirical decompositions of these data.

In the remainder of this chapter, we present the model in several stages. We comment on the choice of degrees of freedom for all the work underlying this chapter. We then introduce the DMFI using η as its primary metric. This is followed by a discussion regarding the choice of LJ parameters. Next, we

introduce the polar components of the model, consisting of a modified short-range electrostatics model and the description of screening of interactions between partial charges due to the local environment. We conclude the presentation of the model by commenting on miscellaneous issues including the treatment of ionic groups and computational efficiency. After sketching the simulation design for the work underlying the results in this paper, we provide a brief history of the calibration of the model. We then present a representative set of preliminary results obtained using ABSINTH. In discussing these results, we attempt to make direct connections with experimental data. We conclude with a brief summary of merits and future improvements to our model.

III.3. The ABSINTH Model

III.3.1. Overview

In ABSINTH, a polypeptide chain is parsed into a series of model compounds corresponding to individual backbone units and sidechains. This is done for the purpose of calculating the DMFI. The sampled degrees of freedom are the dihedral angles and rigid-body coordinates of the macromolecules of interest while bond angles and lengths are held fixed. The ABSINTH Hamiltonian can be written as a sum of the following terms:

$$E_{\text{total}} = W_{\text{solv}} + U_{\text{LJ}} + W_{\text{el}} + U_{\text{corr}} \qquad (3\text{-}1)$$

In Equation 3-1, W_{solv} is the solvation term corresponding to the DMFI. U_{LJ} represents the contributions from short-range steric and dispersive interactions, which are accounted for by the Lennard-Jones model. W_{el} encompasses the

electrostatic model we employ. It is written as W_{el} instead of U_{el} because the mean-field dielectric modulates the interactions based on the conformation of the macromolecule. Finally, U_{corr} represents torsional correction terms applied only to dihedral angles subject to electronic effects, *i.e*., those that cannot be captured by U_{LJ}. In the following paragraphs, all of the terms are explained in detail in the order they appear in Equation 3-1.

III.3.2. Degrees of Freedom

In all of our simulations of polypeptide chains, the degrees of freedom are the backbone and sidechain torsion angles, *viz.* the set of ϕ,ψ, ω, and χ-angles. All bond lengths and bond angles are held fixed. The assumption of fixed bond lengths and angles has been made repeatedly in the literature, and it has been shown recently that in MC simulations such a treatment does not introduce artifacts,[31] unlike in molecular dynamics.[32] However, such constraints can suppress fluctuations necessary for the interconversion between adjacent basins in phase space[33] because the precise nature of constraints is important if one is interested in the quantitative details of barriers, as has been shown in a recent study employing a quantum mechanical Hamiltonian.[34]

III.3.3. Direct Interaction of Solutes with the Mean-Field

The following paragraphs will describe the direct interaction of solutes with the mean-field, *i.e.*, the work done when inserting any solute from vacuum into the continuum solvent while not considering intramolecular terms.[35]

When inserting a rigid molecule into water, there are at least three distinct terms that contribute to the solvation process and the transfer free energy:

1) The purely entropic, unfavorable free energy to create the solute-sized cavity in the dense fluid (cavitation term which is non-polar)[17]
2) The favorable free energy gained from uniform dispersive interactions of the solute with the surrounding water molecules (contributes to the non-polar term)[36]
3) The favorable free energy gained by specific polar interactions of the solute with surrounding water molecules through dipole-dipole or charge-dipole interactions (polar term)[37]

These terms are accounted for by the first few solvation shells.[38] For a rigid solute, our model treats the above three terms "in one shot", *i.e.*, we do not use a formal decomposition.

The use of reference free energies of solvation at the model compound level

We parse the solute into a series of solvation groups, which are all analogs of small, usually rigid model compounds. As an example, the atoms N, H, C, and O of the peptide backbone form a solvation group, and the analog is *N*-Methylacetamide. Figure 3.1 illustrates how we parse the peptide sequence of Met-Enkephalin into solvation groups. For each solvation group, our approach guarantees accurate solvation free energies. This is achieved by construction since for each solvation group we use experimentally measured free energies of solvation presented in Table 3.a.

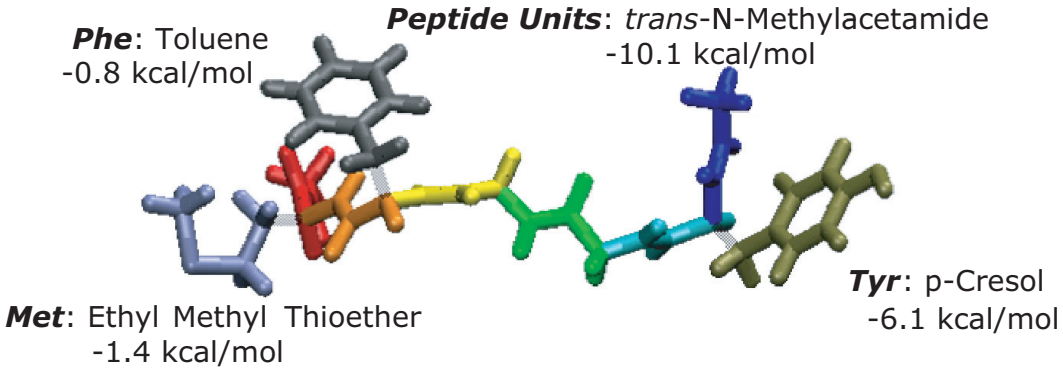

Figure 3.1: **Parsing a solute into model compounds using Met-Enkephalin (Acetyl-YGGFM-N-Methylamide) as an example**. The six peptide units are shown in blue, cyan, green, yellow, orange and red, each using N-Methylacetamide as the model compound. The sidechains for the tyrosine, phenylalanine, and methionine residues are as indicated. The corresponding model compounds are p-Cresol (Tyr), Toluene (Phe), and Ethyl Methyl Thioether (Met). Details of the parsing are shown in Table 3.a.

Residue or Unit	**Model Compound**	**List of Atoms in Solvation Group**	**ΔG_{sol} (kcal/mol) used in ABSINTH**
Polypeptide backbone	N-Methylacetamide	-CO-NH-	-10.1
Formylated peptide N-Cap	N-Methylformamide	-CO-NH-	-10.0
Amidated peptide C-Cap	Acetamide	-CO-NH_2	-9.7
Charged N-terminus	Methylamine	-NH_3	-106.5
Charged C-terminus	Acetate	-COO	-107.3

Residue or Unit	Model Compound	List of Atoms in Solvation Group	ΔG_{sol} (kcal/mol) used in ABSINTH
Glycine	-	-	-
Alanine	Methane	All	+1.9
Valine	Propane	All	+2.0
Leucine	2-Methylpropane	All	+2.3
Isoleucine	Butane	All	+2.2
Proline	Propane	All	+2.0
Methionine	Ethyl Methyl Thioether	-S- $-CH_2-CH_3$, $-CH_3$	-3.6 (Ethyl Methyl Thioether – Butane) +2.2 (Butane)
Serine	Methanol	-OH	-5.1
Threonine	Ethanol	-OH $-CH_3$	-5.1 (MetOH) +0.1 (EtOH-MetOH)
Cysteine	Methanethiol	-SH	-1.2
Asparagine	Acetamide	$-CO-NH_2$	-9.7
Glutamine	Propionamide	$-CO-NH_2$ $-CH_2-$	-9.7 (Acetamide) +0.4 (Propionamide – Acetamide)
Phenylalanine	Toluene	All	-0.8
Tyrosine	p-Cresol	-OH Rest	-5.3 (p-Cresol – Toluene) -0.8 (Toluene)
Tryptophan	3-Methylindole	-NH Rest	-3.5 (3-Methylindole - Naphthalene) -2.4 (Naphthalene)
Histidine	4-Methylimidazole	-NH-C-N-	-10.3

Residue or Unit	Model Compound	List of Atoms in Solvation Group	ΔG_{sol} (kcal/mol) used in ABSINTH
Aspartate (-)	Acetic Acid	-COO	-107.3
Glutamate (-)	Propionic Acid	-COO	-107.3
Lysine (+)	1-Butylamine	$-NH_3$	-100.9
Arginine (+)	n-Propylguanidine	Guanidino Group	-100.9
Sodium (+)	-	Na^+	-87.2
Chloride (-)	-	Cl^-	-74.6

Table 3.a: **Detailed inventory of the solvation groups in ABSINTH**. In general, amino acid residues are partitioned into a sidechain model compound as well as a (universal) backbone model compound. The first column lists the residue name (for specific amino acids referring to sidechains only), the second column gives the model compound used, and the third column lists the atoms making up the solvation group. Note that atoms not listed play no role in the DMFI for that particular residue. Such a choice is often motivated by a single moiety dominating the free energy of solvation. The fourth column lists the reference free energies of solvation as taken from various experimental papers summarized in Marten *et al.*;[39] most prominently the work of Wolfenden[40,41] for net neutral peptide model compounds, and Pliego Jr. and Riveros[42] for ions. Unfortunately, experimental uncertainties are not provided in those original publications. We treat model compounds with distinct polar solvation sites and a significant hydrophobic portion as follows: using the tyrosine sidechain as an example, the difference between the model compound's total free energy of solvation and the underlying hydrophobic model compound (the difference between p-Cresol, -6.1kcal/mol, and toluene, -0.8kcal/mol) is assigned to the hydrophilic functional group(s) (-5.3kcal/mol) while the value for the hydrophobic compound (-0.8kcal/mol) is assigned to the hydrophobic part(s). Within

each subgroup the weight factors for all atoms are uniform. The treatment for isotropic compounds is much simpler (entry “All” in third column). It must be pointed out that the sensitivity to these choices is generally small due to the correlation between the solvation states of the atoms comprising the solvation group. This correlation is also what justifies dropping atoms from the DMFI calculations entirely. The values for charged peptide moieties are lowered artificially by ~30kcal/mol and this was the result of a systematic calibration process (see text).

While the treatment is trivially correct for isolated model compounds, we postulate a model by which the degree of solvent accessibility in larger molecules controls the modulation of the DMFI. This modulation is assessed by evaluating the average solvation state (defined below) for all the atoms comprising the particular solvation group:

$$W_{\text{solv}} = \sum_{i=i}^{N_{\text{SG}}} \zeta_i \cdot \Delta G^i_{\text{solv}} = \sum_{i=i}^{N_{\text{SG}}} \left[\sum_{k=1}^{n_i} \lambda^i_k \cdot \upsilon^i_k \right] \cdot \Delta G^i_{\text{solv}} \qquad (3\text{-}2)$$

In Equation 3-2, N_{SG} is the number of solvation groups in the system, ΔG^i_{solv} is the reference free energy of solvation for solvation group i, and n_i is the number of atoms belonging to solvation group i. The λ^i_k are weight factors ($0 \le \lambda^i_k \le 1$) for the k^{th} atom in solvation group i and the υ^i_k are the corresponding solvation states for individual atoms as discussed below. The choices for the atoms comprising the various solvation groups and their weight factors (λ^i_k) are summarized in Table 3.a and illustrated in Figure 3.1. The λ^i_k are uniform over each subgroup listed in the third column of Table 3.a.

Calculation of atomic solvation states $\left(\upsilon_k^i\right)$ *for the* k^{th} *atom in solvation group* i

The atoms within a solvation group i can be fully solvated ($\upsilon_k^i = 1$), fully desolvated ($\upsilon_k^i = 0$), or partially (de)solvated ($0 \le \upsilon_k^i \le 1$). The latter two states are realized when solvation by water is replaced with solvation by different species. For example, groups buried on the inside of a protein are no longer solvated by water but by the protein core. In order to compute the solvation state for an individual atom, we need to assess the interface of solutes with the surrounding mean-field, *i.e.*, the atomic solvent-accessibilities. These are defined as η_k^i which are the resulting fractions of free volume around an atom k (in solvation group i) after subtracting the atomic volumes of other solute atoms from the maximum accessible volume ($V_{k,\max}^i$), which is defined by the radius of the mean-field solvation shell (see Figure 3.2):

$$V_{k,\max}^i = \frac{4\pi}{3}\left[\left(r_{\mathrm{w}} + \frac{d_k^i}{2}\right)^3 - \left(\frac{d_k^i}{2}\right)^3\right]$$
$$\eta_k^i = 1.0 - \frac{1}{V_{k,\max}^i}\sum_{j=1}^{N_{\mathrm{SG}}}\sum_{l=1}^{n_j}\gamma_{kl}\frac{4\pi}{3}\left(\frac{d_l^j}{2}\right)^3 \qquad (3\text{-}3)$$

Here, r_{w} is the radius of the solvation shell, d_k^i denotes to the diameter of atom k in solvation group i (usually derived from Lennard-Jones parameters, see below), and γ_{kl} is the overlap factor for the solvation shell of atom k with the volume of atom l:

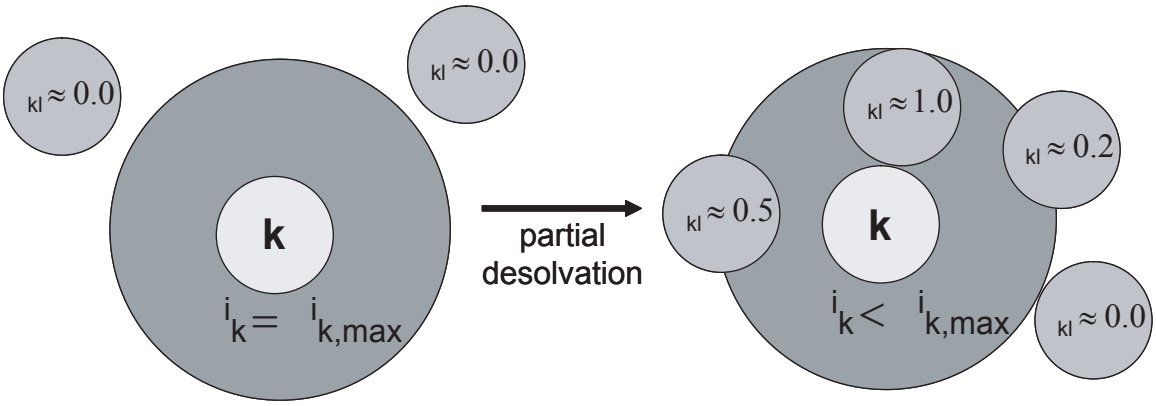

Figure 3.2: **Schematic illustration of the computation of the solvent accessible volume fraction for atom *k* in solvation group *i*,** η_k^i . The light gray circle depicts atom k and the dark gray circle around it its mean-field solvation shell. The medium gray circles indicate other atoms either too far away to affect the solvation of atom k (left side) or occupying part of atom k's solvation shell and consequently reducing η_k^i according to Equation 3-3 (right side).

The solvation state, υ_k^i , will be defined as a function of η_k^i. (also see Figure 3.3). As is clear from Equation 3-3, the η_k^i for a given site can be obtained using the size of the solvation shell (r_w) and the hard-sphere radii of other atoms alone.

To define the fully desolvated state we consider the packing of hard spheres, for which the available space will never be fully used but instead an interstitial space of ~26% will remain. Therefore, if $\eta_k^i \le 0.26$, then atom k in solvation group i is assumed to be fully desolvated, *i.e.*, $\upsilon_k^i = 0$ (see Panel A in Figure 3.3). Conversely, atoms in solvation groups are covalently connected to each other and therefore the upper limit for η_k^i , *viz.* $\eta_{k,\max}^i$ will not be unity. This is because connected atoms will always diminish the accessible volume. To

account for this topology-derived deviation, we adjust the determination of the solvation state of individual atoms to reflect the fact that there is a reduced maximum η_k^i and define this to correspond to $\upsilon_k^i = 1$ (see Panel A in Figure 3.3).

The simplest representation for partially solvated states is shown in Panel A of Figure 3.3, where υ_k^i is a linear function of η_k^i. Instead of a fixed model, one can generalize the interpolation function to be a stretched sigmoid, which provides flexibility in describing the physics of partial desolvation:

$$\upsilon_k^i = \left[1.0 + \exp\left(\frac{-\left(\eta_k^i - d_1\right)}{\tau_d}\right)\right]^{-1} d_2 + d_3$$

$$d_1 = \chi_d \eta_{k,\max}^i + \left(1.0 - \chi_d\right)\eta_{k,\min}^i$$

$$d_2 = \left(\left[1.0 + \exp\left(\frac{-\left(\eta_{k,\max}^i - d_1\right)}{\tau_d}\right)\right]^{-1} - \left[1.0 + \exp\left(\frac{-\left(\eta_{k,\min}^i - d_1\right)}{\tau_d}\right)\right]^{-1}\right)^{-1} \qquad (3\text{-}4)$$

$$d_3 = 1.0 - d_2 \cdot \left[1.0 + \exp\left(\frac{-\left(\eta_{k,\max}^i - d_1\right)}{\tau_d}\right)\right]^{-1}$$

In Equation 3-4, the $\eta_{k,\min}^i$ and $\eta_{k,\max}^i$ are the minimum and maximum expected solvent-accessible volume fractions, which are fixed for a given atom. τ_d is the steepness of the stretched sigmoidal function, and χ_d is its mid-point relative to the limits $\eta_{k,\min}^i$ and $\eta_{k,\max}^i$, respectively.

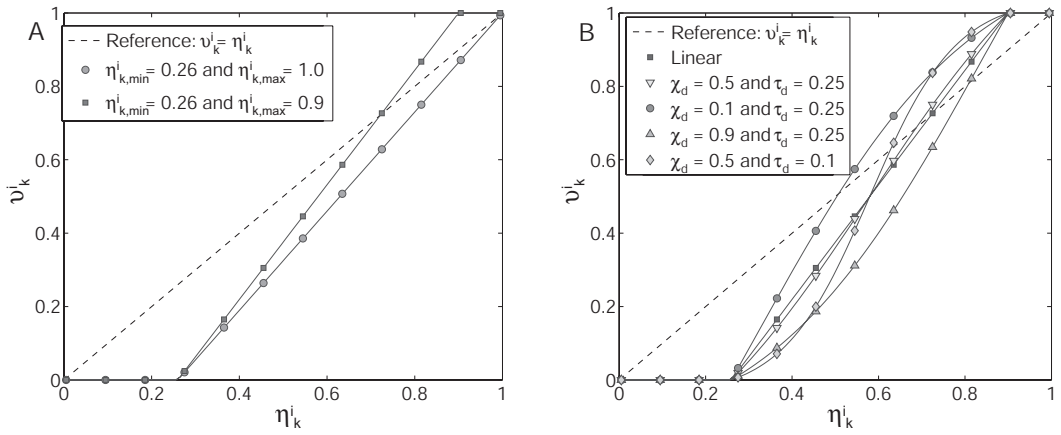

Figure 3.3: **The mapping from the solvent accessible volume fraction η_k^i to the solvation state** υ_k^i . In Panel A, the naïve choice $\upsilon_k^i = \eta_k^i$ is shown along with corrections introduced by the natural bounds of η_k^i (see text). In Panel B, the generalized, sigmoidal interpolation is shown. At τ_d=0.25 and χ_d=0.5, the curve is very similar to the linear case using the same bounds. Shifting χ_d to 0.1 and 0.9, respectively, shifts the mid-point of the transition accordingly but leaves the overall curvature largely unaffected. Conversely, values of χ_d=0.5 and τ_d=0.1 increase curvature and yield a more step-like transition. See Equation 3-4 for details.

For the functional form in Equation 3-4, linear interpolation is recovered in the limit of $\tau_d \rightarrow \infty$, which is true irrespective of the value for χ_d. Conversely, a step function at position χ_d relative to $\eta_{k,\min}^i$ and $\eta_{k,\max}^i$ is obtained in the limit $\tau_d \rightarrow 0$. One might encounter rare cases where η_k^i falls below $\eta_{k,\min}^i$ or exceeds $\eta_{k,\max}^i$. In such cases, the solvation state is set to be zero or unity, respectively. Panel B of Figure 3.3 shows how τ_d and χ_d control the variation of υ_k^i as a function of η_k^i .

The choice of particular values for τ_d and χ_d defines the response of the system to a physical perturbation in which water molecules either enter or exit the hydration environment of a solvated site. Unfortunately, there are no experimental data to help us make the right choices for τ_d and χ_d, respectively. In the absence of such guidance, it seems safe to assume that the linear limit is physically reasonable based on the comparable linearity found for the binding enthalpy of solute-water clusters as a function of the number of water molecules in the clusters.[43] Additionally, hydration numbers are known to be linearly correlated with the magnitude of the solvent interface.[44]

Summary of the DMFI

Polypeptide chains are decomposed into solvation groups which are analogs of model compounds (see Figure 3.1 and Table 3.a). Similar to EEF1 but unlike GB and PB models, the polar and non-polar parts of the transfer process are treated simultaneously using reference free energies of solvation for the solvation groups. Compared to EEF1, we use a different way to measure solvent accessibilities which are fed into a generalized, stretched sigmoidal function to compute solvation states. Finally, we choose model compounds as solvation groups, which allows us to use experimental data for their free energies of solvation directly.

All continuum models of solvation have to provide a quantitative description of partially solvated states. For example, in both GB and PB models, the definition of the dielectric boundary will influence the estimate of charging

free energies. In PB, this estimate will be particularly sensitive to the surface description of the dielectric boundary in regions with high curvature,[45-47] whereas in GB this sensitivity is manifest in the model chosen to calculate the effective Born radii.[13,25,48-52]

III.3.4. Treatment of Steric and Dispersive Interactions

We employ the commonly used Lennard-Jones 12/6-potential to describe both steric repulsions and the weak dispersive attractive interactions:

$$U_{LJ} = 4\sum_{i}\sum_{j>i} f_{ij}\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \qquad (3\text{-}5)$$

In Equation 3-5, r_{ij} is the distance between atoms i and j, ε_{ij} are the pairwise dispersion parameters, and the σ_{ij} are the pairwise size parameters. f_{ij} is unity for pairs of atoms separated by at least one rotatable bond and zero otherwise. The ε_{ij} and σ_{ij} are obtained from the ε_{ii} and σ_{ii} through geometric and arithmetic combination rules, respectively. The choices for the ε_{ii} and σ_{ii} are adaptations from Pauling / Hopfinger's values which were parameterized to reproduce physical properties of small molecule crystals.[53] The choices for the σ_{ii} differ considerably from values used in classical force fields. These differences are motivated based on the following considerations:

In most classical force fields, physical data for neat liquids, most notably densities and heats of vaporization, are used to fit LJ parameters for the atom types occurring in the small molecules comprising the calibration set.[54,55] By necessity, however, these parameters will be co-dependent on the set of partial

charges employed, which immediately questions their transferability, in particular to a continuum solvation model.[5,10]

The transferability can be questioned in terms of the size parameters since the concatenation of small molecules into polymers generates new torsional degrees of freedom for which the rotational barriers will usually have to be corrected by applying elaborate torsional potentials. We do not have to employ these correction terms since the size parameters we employ are substantially smaller than those in standard force fields. We have shown that a variant of our LJ parameters gives an accurate account of local steric effects in polypeptide chains.[56] Moreover, the transferability can be questioned in terms of the interactions strengths because the hydrophobicity with respect to a given water model will not have been calibrated properly. The appropriate test for the latter is to computationally determine the transfer free energies for these small molecules from vacuum into water. Such studies[57-60] have usually revealed some systematic flaws in the traditional force fields and have primarily been used to improve the charge sets employed.[61,62] Interestingly, it has been noted that it might be impossible to unify both sets of calibration data, *i.e.*, both neat liquid data as well as transfer free energies, with a single set of fixed-charge parameters.[54,59,60,62] However, the steric and dispersive parameters are usually excluded from these improvements. Hence, we use LJ parameters which are chemically accurate rather than the result of a fitting procedure that requires us to rely on the assumption of transferability. They are summarized in Table 3.b:

Atom Type	Example	σ_{ii} in Å	ε_{ii} in kcal/mol	Valency
Aliphatic or aromatic N (sp^2)	Amide N	2.70	0.150	3
Aliphatic N (sp^3)	Amine N	2.70	0.150	4
Non-protonated, aromatic N (sp^2)	Imidazole N	3.20	0.150	2
Proline N (sp^2)	Proline N	2.70	0.150	3
O (sp)	Carbonyl O	2.70	0.200	1
O (sp^2)	Alcohol O	3.00	0.150	2
Aliphatic C (sp^3)	Methyl C	3.30	0.100	4
Aromatic or aliphatic C (sp^2)	Phenyl C	3.00	0.100	3
Non-polar H	Methyl H	2.00	0.025	1
Polar H	Alcohol H	2.00	0.025	1
Na^+	Sodium Ion	3.33	0.003	0
Cl^-	Chloride Ion	4.42	0.118	0

Table 3.b: **Summary of Lennard-Jones parameters**. These parameters were used for most of the results presented in this and subsequent chapters. The first column lists atom types with hybridization states, the second column provides a chemical example for every atom type, the third and fourth columns list the actual LJ parameters σ_{ii} and ε_{ii}, and the fifth column gives the valency of each atom type. Ion parameters are loosely based on the Åqvist parameters in the OPLS-AA force field.

III.3.5. Treatment of Polar Interactions

Polar interactions are typically viewed as the primary determinant of specificity in biomolecular interactions. In almost all classical force fields intended

to work with explicit water models they are treated by applying Coulomb's law to the interactions of a set of carefully determined, fixed point charges.

Short-range electrostatics in the point-charge approximation

A majority of functional groups in polypeptides are polar and net-neutral. Dipole moments of these functional groups are modeled using point charges. Therefore, a majority of electrostatic interactions involve groups of point charges that are net-neutral, and interactions should only be evaluated between those charge groups. Violation of this rule leads to the computation of spurious charge-dipole and charge-charge interactions although the charges will be fractional. This issue arises for atoms which are close due to chain connectivity since bonded interactions (separated by one (1-2) or two bonds (1-3)) are excluded from the non-bonded energy calculation. Classical force field development has addressed this problem through the use of torsional potentials as well as *ad hoc* factors to scale interactions between atoms separated by three bonds (1-4).

A recent study has shown that the manipulation of these *ad hoc* factors can impact the predictions made by force fields even in simulations using explicit solvent.[8] In many implicit solvent calculations, however, the presence of many-body terms will overemphasize the effects of ill-represented short-range interactions. To circumvent this problem, we re-formulate the electrostatic model. We only include interactions between net-neutral groups of point charges, unless the functional group has a net charge. These groups will collectively be referred to as charge groups. Consequently, the electrostatic interactions are written as:

$$W_{\mathrm{el}} = \sum_{i=1}^{N_{\mathrm{CG}}} \sum_{k=1}^{n_i} \sum_{j=i+1}^{N_{\mathrm{CG}}} \sum_{l=1}^{n_j} f_{ij} \frac{q_k^i q_l^j}{4\pi\varepsilon_o r_{kl}} s_{kl} \qquad (3\text{-}6)$$

In Equation 3-6, N_{CG} is the number of charge groups in the system, $n_{i(j)}$ is the number of point charges in charge group $i(j)$, and the q_k^i and q_l^j are the charges on the k^{th} and l^{th} atom in charge groups i and j, respectively. r_{kl} is distance between atoms k and l and s_{kl} denotes the net screening factor (see below). ε_0 is the vacuum permittivity, and f_{ij} is a factor which assumes a value of zero if charge groups i and j possess any pair of atoms k and l that are (1-2)- or (1-3)-bonded to one another. Otherwise, f_{ij} assumes a value of unity. The functional form implies that there can never be any polar interactions within a charge group. Additionally, interacting charge groups cannot have any pair of atoms separated by less than a single rotatable bond. This modification has no major consequences on the majority of the polar interactions because they are largely non-local.

For a given polypeptide, the number and composition of the charge groups will depend on the charge set, *i.e.*, the molecular mechanics force field from which we obtain the charges. Our model is best-suited for charge sets such as OPLS-AA[55] or GROMOS[54] in which charge groups are typically small and localized. Conversely, charge sets such as AMBER[63] or CHARMM[64] with significant pre-polarization in the fixed charges seem less well-suited. This is due to their large charge groups, which would result in the complete elimination of local polar interactions. We will present results from tests on different charge sets. As was noted previously, charge sets in classical force fields are co-

parameterized along with LJ and other parameters, although the extent of co-parameterization depends on the specific paradigm adopted by a force field. Consequently, it might seem counterintuitive to treat the LJ and charge parameters as modular entities. We believe that rigorous co-dependence of parameters is valid only in the limit of neat liquids or dilute binary mixtures of small molecules in aqueous solution. Beyond this regime, numerous approximations and assumptions are required to transfer model compound parameters for use in simulations of polypeptides. Additionally, the use of similar parameter sets for simulations with explicit versus implicit solvation models has been questioned in general.[5,10] Therefore, we see no *a priori* reason to maintain strict adherence to the coupling paradigm adopted by a specific force field. Instead, we converged on the modular approach of using Pauling-style LJ parameters and allowing flexibility in the choice of charge sets. For the work in this and the following chapters we primarily use the OPLS-AA/L[65] charge set because it fits well with our approach for modeling electrostatic interactions (see Equations 3-6 and 3-9).

Solvent-modulation of Coulombic interactions

The remaining component of the model is the screening of Coulombic interactions by the continuum dielectric. In PB/GB models, screened Coulombic interactions are coupled to the polar component of the transfer process. In the GB formalism,[13] the polar contribution to the solvation free energy is written as follows:

$$G_{\mathrm{pol}} = -\frac{1}{2}\cdot\left(1-\frac{1}{\varepsilon_{\mathrm{w}}}\right)\sum_{i=1}^{n}\sum_{j=i}^{n}\frac{q_i q_j}{f_{\mathrm{GB}}}$$
$$f_{\mathrm{GB}} = \left[r_{ij}^2 + \alpha_i\alpha_j \exp\left(\frac{-r_{ij}^2}{4\alpha_i\alpha_j}\right)\right]^{0.5} \qquad (3\text{-}7)$$

Here, ε_{w} is the dielectric constant of water, $q_{i(j)}$ denotes the charges on atoms $i(j)$, r_{ij} is the distance between the two atoms, and the α_i and α_j are the generalized Born radii for atoms i and j, respectively. While the sum can formally be decomposed, the screening process and the polar component of the DMFI remain coupled through the Born radii as shown below:

$$G_{\mathrm{pol}} = -\left(1-\frac{1}{\varepsilon_{\mathrm{w}}}\right)\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{q_i q_j}{f_{\mathrm{GB}}} - \frac{1}{2}\cdot\left(1-\frac{1}{\varepsilon_{\mathrm{w}}}\right)\sum_{i}^{n}\frac{q_i^2}{\alpha_i} \qquad (3\text{-}8)$$

In the cross-term (first term on the right-hand side of Equation 3-8), the Born radii de-screen polar interactions between buried charges since those will have large values for the α_i.

In ABSINTH, we handle the transfer process separately. Therefore, only the modulation of solute-solute polar interactions needs to be dealt with at this stage. In ABSINTH, the solvation states υ_k^i replace the Born radii as indicators of how buried or solvent-accessible the charges are, and the total Coulomb energy is written as:

$$W_{\mathrm{el}} = \sum_{i=1}^{N_{\mathrm{CG}}} \sum_{k=1}^{n_i} \sum_{j=i+1}^{N_{\mathrm{CG}}} \sum_{l=1}^{n_j} f_{ij} \left[1 - a\upsilon_k^i\right]\left[1 - a\upsilon_l^j\right] \frac{q_k^i q_l^j}{4\pi\varepsilon_{\mathrm{o}} r_{kl}}$$

$$= \sum_{i=1}^{N_{\mathrm{CG}}} \sum_{k=1}^{n_i} \sum_{j=i+1}^{N_{\mathrm{CG}}} \sum_{l=1}^{n_j} f_{ij} \frac{q_k^i q_l^j}{4\pi\varepsilon_{\mathrm{o}} r_{kl}} - \sum_{i=1}^{N_{\mathrm{CG}}} \sum_{k=1}^{n_i} \sum_{j=i+1}^{N_{\mathrm{CG}}} \sum_{l=1}^{n_j} f_{ij} \left[a\left(\upsilon_k^i + \upsilon_j^l\right) - a^2 \upsilon_k^i \upsilon_l^j\right] \frac{q_k^i q_l^j}{4\pi\varepsilon_{\mathrm{o}} r_{kl}} \quad (3\text{-}9)$$

$$a = \left(1.0 - \frac{1}{\sqrt{\varepsilon_{\mathrm{w}}}}\right)$$

The product of the two square brackets in the first line of Equation 3-9 is the screening factor, s_{kl}, for this interaction (see Equation 3-6). Note that Equation 3-9 corresponds to the first term in Equation 3-8. In Equation 3-9, there is no term corresponding to the second one in Equation 3-8 since the polar part of the DMFI is an integral part of the free energies of solvation (Equation 3-2).

The use of solvation states ($\upsilon_k^i, \upsilon_l^j$) in both the Coulombic screening (Equation 3-9) and the DMFI (Equation 3-2) would allow us to couple these two processes. However, such models have only two adjustable parameters. Initial tuning indicated that when the two terms are coupled the free energy of solvation term dominates and therefore conformations that are maximally solvated are generally preferred (data not shown). Therefore, we define a second stretched sigmoid analogous to the one in Equation 3-4 to determine the solvation states, υ_k^i and υ_l^j, for use in Equation 3-9. For the second function, the parameters χ_{d} and τ_{d} are replaced with different parameters χ_s and τ_s, respectively. If $\chi_{\mathrm{s}}=\chi_{\mathrm{d}}$ and $\tau_{\mathrm{s}}=\tau_{\mathrm{d}}$, then the values for υ_k^i in Equations 3-2 and 3-9 are identical. The physical reason for using independent parameters is the different nature of the two processes described. We cannot assume that the free energy contribution from

the DMFI responds to changing numbers in water molecules in the hydration shell in the same way as the dielectric response which leads to screening of polar interactions. This decoupling is similar to the use of different interfaces in PB/GB models for non-polar versus polar components, because the dielectric boundary does not necessarily coincide with the surface definition used to determine the non-polar contribution to the solvation energy.

To summarize the foregoing discussion, the central difference between the ABSINTH / EEF1 paradigm and the PB/GB paradigm lies in the treatment of the solvation process. In the former, the DMFI comprising both polar and non-polar contributions is considered “in one shot” but the screening of polar interactions by the continuum dielectric has to be considered separately. Conversely, in PB/GB models, the polar part of the DMFI and the dielectric screening are coupled and captured “in one shot”. Here, the non-polar contributions to the solvation process have to be considered independently.

III.3.6. Miscellaneous

Using specific torsional potentials to restrain pseudo-rigid bonds

By omitting torsional potentials, we prescribe that the majority of rotational barriers can be captured by excluded volume interactions. However, there are certain cases where electronic effects lead to strong rotational barriers, and we handle these separately. The amide bonds along the peptide backbone are quasi-rigid, and we employ torsional potentials taken directly from the OPLS-AA force field[55] to keep the peptide dihedral (ω) predominantly in the *trans*-

configuration. It has been argued that oscillations of the ω-angle mediate crucial correlations between the surrounding dihedral angles;[66] this supports the view that constraining these degrees of freedom might suppress conformational flexibility. Similarly, we adopt torsional potentials for the rotation of the polar hydrogen in the tyrosine sidechain which – against steric preferences - favors an in-plane arrangement.

The treatment of ionic groups

In principle, the paradigm outlined so far may be applied to solvation and charge groups carrying a net charge as well, such as mobile counterions or charged moieties in polypeptides. The solvation properties of ionic groups pose unique challenges for all continuum electrostatic models.[67] There are several reasons for this but in general one can argue that dipolar and ionic solvation differ fundamentally from each other, as is evidenced by the large body of theoretical and experimental work dedicated exclusively to electrolyte solutions.[68]

An obvious advantage of the ABSINTH paradigm is that inorganic ions are represented explicitly. This means that correlations due to finite size are addressed automatically. In this sense, the model is similar to extensions of PB theory, which add explicitly represented counterions.[69] The LJ and free energy of solvation parameters used to model these ions in the bulk are listed in Tables 3.b and 3.a.

Special consideration is required for treating ionic groups that are part of the polypeptide chain. Free energies of solvation for monovalent, organic or

inorganic ions typically range from -50 to -100kcal/mol[42,70] and are an order of magnitude larger than the values for neutral, small molecules. Nonetheless, desolvation of charged moieties in polypeptides might be favorable due to electrostatic interactions of equivalent strength, such as salt bridges. Due to the large magnitude of the energies, the balance between these two effects is very sensitive if the same paradigm (Equations 3-2 and 3-9) is used for ionic solvation as is for dipolar solvation. If the balance tips over to the desolvated side, the system can become trapped in deep, local minima; either because the mean-field nature of the model and the finite sampling suppress the necessary fluctuations to escape from such minima or because they are in fact stable states for the particular Hamiltonian. Due to recurring problems with desolvated charges (data not shown), we lowered the values used for the free energies of solvation of charged peptide moieties substantially (see Table 3.a) while maintaining an identical paradigm (Equations 3-2 and 3-9) for all solvation groups in the system. The only other modification vis-à-vis electrostatic interactions between neutral moieties is that we ignore cutoffs for groups carrying a net charge (in reference to Equation 3-6).

Computational Efficiency

The model including the DMFI but excluding the screening of polar interactions is as efficient as gas phase calculations using the same underlying non-bonded potential functions. This is possible because we compute solvation states of individual atoms using the same distance information required to

compute short-range, non-bonded interactions given certain simplifying assumptions. These assumptions are as follows:

1) We treat all atoms as spheres with a well-defined radius.
2) Spherical envelopes of covalently bound atoms will overlap and hence we use a pre-computed, pairwise correction term to reduce the volume of such atoms by subtraction.
3) We use linear approximations to assess all spherical overlaps. These work reasonably well providing the radii of the spheres are roughly comparable.
4) Overlaps involving three or more spheres are assumed to be negligible.[28]

While more complicated expressions could be used,[71] the qualitative nature of the model and the goal to be as efficient as possible justify the simpler choice.

The screening of polar interactions poses more of a challenge, as effective three-body interactions become possible, *i.e.*, the Coulomb interaction between two (partial) charges is in fact a function of the coordinates of other nearby atoms due to their effect on the solvation state of the two charges. For MC simulations, this implies that upon a proposed move more energy terms need to be evaluated than just the ones involving atoms that moved relative to one another. We have implemented a detailed bookkeeping scheme to track the interactions that change with different MC move sets. This significantly reduces the overhead associated with the computation of screened electrostatic interactions. With these approximations in place, the computational expense for simulations increases by factors of ~2.0-5.0 with respect to gas-phase calculations.

III.4. Simulation Details

This section will provide the details of the simulation setup for the different test systems. All simulations were performed using MC sampling (see Table 3.c) in the canonical ensemble with a spherical droplet boundary condition. The latter was generally modeled using a stiff harmonic potential (see Equation 4-1). The peptides were built according to the Engh-Huber high-resolution, crystallographic geometries,[72] and the sampled degrees of freedom encompassed all rotatable (ϕ,ψ,χ) and some semi-rigid dihedral angles, in particular the peptide ω-angle as well as the χ-angle describing the rotation of the polar hydrogen in tyrosine. All other semi-rigid dihedrals such as those in aromatic rings were held fixed.

	NMR	**Proteins**	**FS**	**Trpzip**	**PolyQ**
Rigid Body	0% / 0% / 1% (90%, 5Å, 60°)	5% (75%, 2.5Å, 25°)	10% (50%, 2.0Å, 10°)	5% (50%, 2.0Å, 10°)	0%
Sidechain (χ_i,χ_j)	0% / 25% / 24.8% (2x, 60%, 30°)	14.3% (2x, 60%, 30°)	9% (2x, 60%, 30°)	28.5% (3x, 60%, 30°)	30% (4x, 60%, 30°)
Pivot (ϕ,ψ)	90% / 67.5% / 66.8% (70%, 10°)	65.2% (70%, 10°)	58.3% (70%, 10°)	47.9% (70%, 10°)	37.8% (70%, 10°)

	NMR	Proteins	FS	Trpzip	PolyQ
Omega (ω)	10% / 7.5% / 7.4% (85%, 5°)	11.5% (85%, 5°)	6.5% (90%, 5°)	5.3% (90%, 5°)	4.2% (90%, 5°)
Concerted Rotations: Four (ϕ,ψ) pairs in concert	0% / 0% / 0%	4%	16.2%	13.3%	28%

Table 3.c: **Overview of the details of the move sets employed for individual systems** discussed in the Results section. The first column lists the degrees of freedom sampled by a particular type of move. Rigid-body moves are always coupled and sample global rotational and translational degrees of freedom. These moves are especially important for the simulations of the two proteins, the FS peptide, and "trpzip1", because the droplet consists of the polypeptide, neutralizing counterions, and excess salt. The concerted rotation approach[73] samples four consecutive sets of backbone ϕ, ψ-angles. The second through fifth columns give the frequencies (in percent) with which the specific move type (row element) is picked for each system. "NMR" stands for coupling constants, "Proteins" refers to the thermal unfolding of two small proteins, "FS" and "Trpzip" indicate the reversible folding of the FS-peptide and trytophane zipper, respectively, and "PolyQ" stands for the polymeric properties of polyglutamine. There are three separate values listed for the coupling constant work, which are for alanine (no χ-angles), net neutral dipeptides, and net charged dipeptides, respectively. Additional information is given in parentheses, indicating what portion of the moves of a certain type consists of stepwise perturbations of the respective degree(s) of freedom, along

with the maximum step size. The remaining fraction consisted of moves fully randomizing the respective degree(s) of freedom. In addition, due to their low computational complexity, sidechain moves consist of multiple identical cycles indicated by the first entry in parentheses.

We used spherical cutoffs of 12.0Å for Coulomb interactions between net-neutral charge groups. No cutoffs were used for interactions involving ionic groups. Cutoffs for the short-range interactions were chosen to ensure maximum accuracy for the computation of the η_k^i and ranged from 9.0-10.5Å for the different simulations. For the results presented here, we used the values shown in Table 3.d for ε_w, r_w, τ_s, χ_s, τ_d, and χ_d, respectively.

r_w in Å	τ_d	χ_d	τ_s	χ_s	ε_w
5.0	0.25	0.1	0.5	0.9	78.2

Table 3.d: **Parameters of the continuum solvation model**, which are used in all ABSINTH calculations presented in this thesis.

We explore different LJ parameters and charge sets in our studies of NMR coupling constants. For all other calculations we choose the OPLS-AA/L charges[55] in conjunction with the LJ parameters shown in Table 3.b. The software used was our in-house MC package (CAMPARI)[74] developed alongside the continuum model presented here.

NMR Coupling Constants

All twenty naturally occurring amino acids except glycine and proline were modeled as dipeptides (Acetyl-X-N-Methylamide) in a droplet of 125.0Å radius

along with a neutralizing counterion (Na^+ or Cl^-) when appropriate. The simulation temperature was 298K and a total number of 2×10^6 MC moves were attempted, while statistics for the coupling constants were accumulated every ten steps. For details of the move set employed, see Table 3.c. For an individual conformation, the coupling constant between the hydrogen atoms at the N- and the C_α-position was calculated using the Karplus relation:[75]

$$^3J\left(H_{N,}H_\alpha\right) = a\cdot\cos^2\phi' - b\cdot\cos\phi' + c \qquad (3\text{-}10)$$

Here, ϕ' is the effective dihedral angle between the two hydrogen atoms of interest, and is directly proportional to the backbone angle ϕ. For the empirical parameters a, b, and c, we use the same strategy as Avbelj and Baldwin in their work on the coil library,[76] *i.e.*, we averaged over four independently obtained sets of these parameters.

Thermal Unfolding of two Small Proteins

The B1 domain of protein G (PDB accession code: 1GB1) and the engrailed homeodomain (PDB accession code: 1ENH) were, after a brief minimization and relaxation to the Engh-Huber geometry, used as starting structures for simulations in a droplet of 75.0Å radius. To reduce the complexity of the calculation while maintaining a somewhat realistic electrolyte environment, the protein was simulated in the presence of neutralizing counterions (the net charges of the proteins are -4 and +7, respectively) and a low-salt background of

either ~9 mM NaCl (1GB1) or ~13 mM NaCl (1ENH). The simulations were carried out at evenly spaced temperatures from 260K to 440K and consisted of 2.5×10^7 MC steps the first 10^7 of which were discarded as equilibration. For calculating the RMSD values, structures were saved every 10^5 steps, while polymeric quantities were averaged every 100 steps. Details of the move sets for all simulations are summarized in Table 3.c.

Reversible Folding / Unfolding of a Helical Peptide

The FS-peptide (Acetyl-$A_5(AAARA)_3$-N-Methylamide) was simulated in a droplet of 45.0Å radius in the presence of neutralizing counterions (the net charge of the peptide is +3) as well as a low-salt background of ~15 mM NaCl. The simulations were carried out at evenly spaced temperatures from 260K to 440K and used either a perfect α-helix (unfolding runs) or random extended conformations (folding runs) as their starting conformations. For details of the move set employed, see Table 3.c.

The data were analyzed according to Lifson-Roig (LR) theory for helix-coil transitions.[77] The α-basin in ϕ,ψ-space was defined as a roughly spherical area around the ideal α-helix geometry with a radius of ~30° largely in agreement with previous work by others.[78,79] Statistics of the backbone angles ϕ and ψ were recorded every ten steps and the distribution of segments with one or more consecutive residues in α-helical conformation was obtained. From this, the LR nucleation and propagation parameters are accessible through a fitting procedure:[78-80]

$$\langle N_h \rangle = \frac{\partial \ln Z}{\partial \ln w}$$
$$\langle N_s \rangle = \frac{\partial \ln Z}{\partial \ln v_{12}}$$
$$Z = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \mathbf{M}^n \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}^T \qquad (3\text{-}11)$$
$$\mathbf{M} = \begin{pmatrix} w & v & 0 \\ 0 & 0 & 1 \\ v & v & 1 \end{pmatrix}$$

Here, $<N_h>$ and $<N_s>$ describe the average number of helical hydrogen bonds and number of helical segments of at least two residues in length, respectively. Z is the partition function in the LR theory and is written in matrix form using the statistical weight matrix $\mathbf{M}$. The latter contains the helix propagation parameter w and the helix-nucleation parameter v, both of which are fit by matching the expected number of helical segments and hydrogen bonds to the computational data using segment statistics. The symbol v_{12} refers to v in the first row and second column of $\mathbf{M}$, *i.e.*, the partial derivative is with respect to that element alone.

Reversible "Folding / Unfolding" of a β-Hairpin Peptide

The peptide SWTWEGNKWTWK-NH_2 was simulated in a droplet of 45.0Å radius in the presence of neutralizing counterions (in accordance with experiment,[81] the N-terminus is modeled as charged bringing the net charge of the peptide to +2) as well as a low-salt background of ~20mM NaCl. The starting structure for the unfolding runs was the NMR structure (Model 1 in 1LE0), which was used after a brief minimization to conform it to the Engh-Huber geometries, and for the folding runs we employed random extended conformations. The

simulations were carried out at evenly spaced temperatures from 260K to 440K and comprised of 4×10^7 MC steps with 2.5×10^7 steps of equilibration. For details of the move set employed, see Table 3.c.

The data were analyzed by computing various orders parameters for 10^4 snapshots for each individual simulation. The RMSD was computed for all heavy backbone atoms excluding the N-terminal serine and the C-terminal amide group. The radius of gyration of the hydrophobic cluster was calculated by taking into account the atoms of the four tryptophan sidechains. An average strand-to-strand distance was defined by computing the average distance between heavy backbone atoms (N, C_α, and C) on one strand and their properly aligned counterparts on the other strand assuming a perfectly symmetrical hairpin. This includes for example atom pairs Glu5:N / Lys8:C or Thr3:C_α / Thr10:C_α. The order parameter L was obtained from Snow *et al.*,[82] and represents the sum of native hydrogen bond distances as well as CD2-CD2 distances for tryptophan sidechains found in contact in the NMR ensemble. Finally, hydrogen bonds were counted if the distance between donor nitrogen and acceptor oxygen atoms on opposite strands was less than 4.0Å.

Polymeric Behavior of Polyglutamine

Acetyl-$(Gln)_N$-N-Methylamide was modeled and simulated for chain lengths of N=20, 24, 27, 33, 36, 40, 47, *i.e.*, for chain lengths mostly in accordance with a recent fluorescence correlation spectroscopy (FCS) study.[6] The simulation system in each case was a droplet with a fixed radius of 130.5Å,

large enough to accommodate fully extended chains. This eliminates all potential boundary artifacts. The simulation temperature was 298K and a total number of $(N/2)\times10^6$ MC moves were attempted for each of the four independent replicas for each chain length (N). The details of the move set are summarized in Table 3.c.

III.5. Calibration of the ABSINTH Model

In this paragraph, we summarize a few of the major steps involved in advancing the model to its current state. The basic paradigm of the model used to describe the DMFI of the solutes with the continuum has provided the relatively rigid framework within which all further development was carried out. Using the "traditional" model – including (1-4)-scaling – for the treatment of short-range electrostatic interactions tended to generate unreasonable results for the conformational preferences of dipeptides, which caused us to design the modified model presented above. We also found that for solutions of small molecules we encountered a lack of favorable intermolecular interactions when using a linear mapping from η_k^i to υ_k^i with the same parameters employed for both the DMFI and the screening of Coulombic interactions. The introduction of both the generalized sigmoidal interpolation function (see Equation 3-4) and the de-coupling of the interpolation parameters χ and τ for the two different aspects of solvation helped eliminate this deficiency with respect to calibration results obtained in explicit solvent. At this juncture, several test simulations on a variety of systems including short peptides, solutions of small model compounds, and

the stabilities of small proteins indicated that the model reproduced expected data reasonably well (based on comparison to data from all-atom molecular dynamics (MD) simulations or to expectation derived from experimental evidence). The remainder of the development then focused on testing various parameter sets for the ε_{ij}, σ_{ij}, and partial charges and on the optimization of the solvation parameters τ_s, χ_s, τ_d, and χ_d.

Work on longer peptides, which show reversible folding, remained largely unsuccessful, until the crucial modification of increasing the size of the solvation shell radius, r_w, from the original value of 2.8Å to 5.0Å. In retrospect, the larger value for r_w is in accord with locations of first hydration shell water molecules around most of the solvation groups used in this work (calibration data not shown). Thereafter, the testing continued by re-assessing the choices for all the parameters, including charge sets and LJ parameters, in the context of results for the reversible folding of α-helix- and β-hairpin-forming peptides. These studies were complemented by continuing work on assessing local steric preferences for peptides (through quantitative comparison of NMR coupling constants) and through work on intrinsically disordered polypeptides, such as polyglutamine.

The preceding summary neglects many of the choices explored during the development phase. We wish to remind the reader that – due to computational infeasibility – we did not perform a systematic search of the entire parameter space, specifically for combinations of r_w, τ_s, χ_s, τ_d, and χ_d. Additionally, we have not been exhaustive in calibrating the model on a large number of systems.

Consequently, the true efficacy of the model can only be adjudicated upon following large-scale calibration exercises, which will require significant investment of computational resources. This is part of ongoing work.

III.6. Results

We present results on several different test systems to assess the validity of the ABSINTH model. These are as follows:

1) NMR coupling constants for dipeptides and comparative analysis of alanine dipeptide
2) The thermal unfolding of two small, stable proteins (1GB1 and 1ENH)
3) The reversible folding / unfolding of the FS-peptide
4) The reversible “folding / unfolding” of the tryptophan zipper “trpzip1”
5) The polymeric behavior of the intrinsically disordered polyglutamine peptides as a function of chain length

Briefly, we use NMR coupling constants to motivate our final choice of LJ parameters. To justify our decision to ignore torsional potentials for a majority of rotatable bonds, we present a comparative analysis of the conformational equilibria of alanine dipeptide to published simulation results. We use the thermal unfolding of the two proteins to show that fully folded proteins with differing folds are stable states for the Hamiltonian presented here and that they exhibit authentic, cooperative unfolding in response to thermal denaturation. We demonstrate the ability to simulate reversible melting using the α-helical FS-

peptide which has been a popular model system for computer simulation. For the tryptophan zipper, we present results indicating that the system reversibly adopts a native-like mean topology at low temperature but that the ABSINTH Hamiltonian fails to predict the specific NMR-determined structure as a stable minimum. Finally, we show that the Hamiltonian provides an accurate description of conformational equilibria for intrinsically disordered polypeptides such as polyglutamine. All of the test systems attempt to make direct contact to experimentally obtained results and strive to define analytic measures most closely related to the experimental measurements.

For a Hamiltonian designed to study IDPs, it is insufficient to present calibration data on the stability of folded proteins or on the accurately reproduced experimental numbers for somewhat unrelated calibration systems such as small model compounds. For simulating self-assembly, it is crucial to describe both the generic polymer character of these macromolecules as well as the stability of the structural preferences they might exhibit. In this light, it seems “safer” to underpredict the latter rather than to follow the approach taken by standard force fields, which commonly overpredict structural preferences, as they are designed to primarily simulate the folded ensembles of polypeptides (see III.2). This is achieved partially through a local pre-organization of the backbone as is demonstrated in the next section.

III.6.1 NMR Coupling Constants and Conformational Equilibria for Alanine Dipeptide

Vicinal, $^3J(H_\alpha,H_N)$, proton-proton coupling constants report primarily on the ϕ-angle of the polypeptide backbone. The relationship between the measured coupling constants and ϕ is expressed via the Karplus equation[75] shown in Equation 3-10. This equation has been parameterized repeatedly to provide better predictive power for structure determination using the $^3J(H_\alpha,H_N)$. It should be pointed out, however, that their extraction from simulation data is non-trivial, in particular in the presence of large conformational fluctuations.[83] This fundamental inability to connect simulation results to experimental readouts provides a potential explanation for some of the discrepancies encountered below. This is particularly true for the systems studied here, *i.e.*, extremely short peptides which will quickly switch from one conformational state to the other. In our analysis, we assume that each snapshot is an independent, "zero-motion" member of the NMR ensemble.

Figure 3.4 shows results using the ABSINTH model coupled to charge and LJ parameters from three common force fields while ignoring all other terms inherent to these force fields, *i.e.*, torsional potentials. Coupling constants obtained from simulation are plotted against the experimental values for dipeptides at pH 4.9[84] along with values obtained through coil library fits for all common amino acids with the exception of glycine and proline. Aspartate, glutamate, lysine, and arginine were modeled in their charged states, while histidine was modeled in its neutral state:

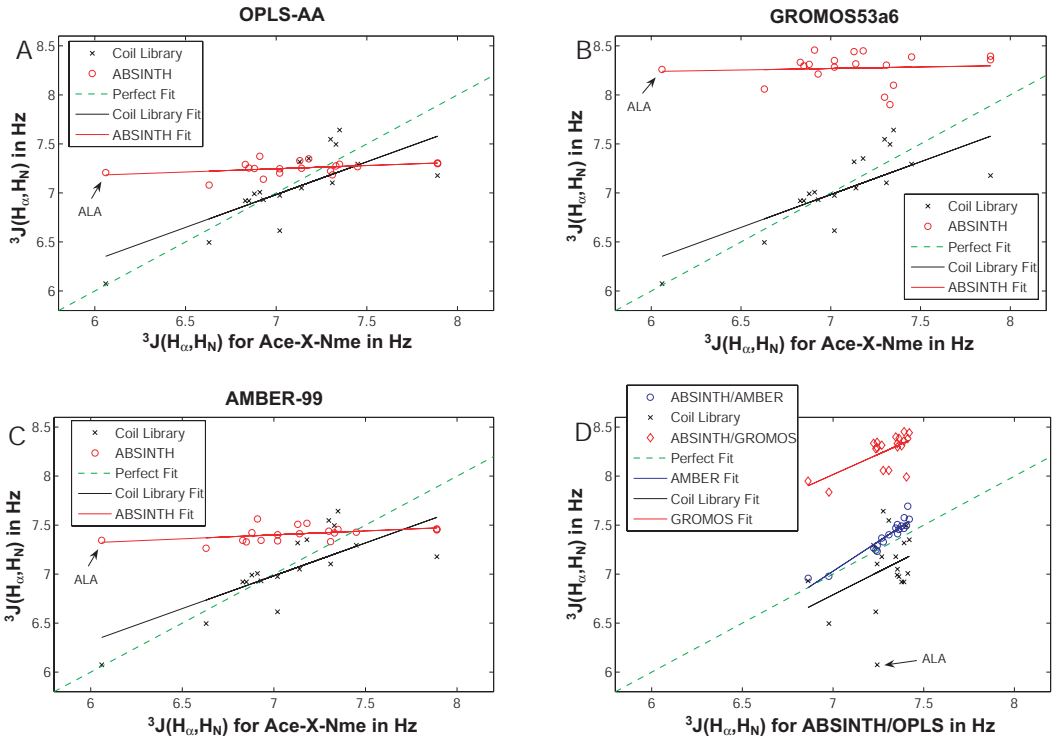

Figure 3.4: **NMR $^3J(H_\alpha N_H)$ coupling constants obtained using ABSINTH's continuum solvation model coupled to standard force field parameters**. Panel A shows the correlation between values measured by Avbelj *et al.* to the coil library as well as ABSINTH/OPLS-AA/L. Panels B and C show analogous plots for ABSINTH/GROMOS and ABSINTH/AMBER, respectively. Finally, Panel D shows a comparison of the values obtained with ABSINTH/OPLS-AA/L to the other two computational models as well as the coil library. Alanine is indicated in all plots as the most drastic outlier.

Panel A of Figure 4 shows that the values obtained for the OPLS-AA/L force field (circles) are insensitive to the type of sidechain, and that they are generally too large when compared to the direct measurements. Alanine is the most drastic outlier as indicated on the plot but the agreement is generally poor.

The values obtained from the coil library fits[76] show better agreement with experiment, although the slope of the correlation is less than unity for both comparisons implying larger similarity between simulated values and coil library fits compared to simulation and (direct) experiment. This suggests that the application of the Karplus equation to extract coupling constants inherently gives rise to some similarity but might always deviate somewhat from direct measurements of the $^3J(H_\alpha,H_N)$.

The situation for the AMBER-99 force field is almost identical (Panel C) even though the values for the coupling constants are slightly larger and hence further away from the measured values. Finally, the GROMOS53a6 force field (Panel B) is unable to generate reasonable coupling constants because aliphatic hydrogen atoms - including the peptide α-hydrogen - are not actually steric interaction sites. This removes an important barrier for the ϕ-angle, normally separating the β- and polyproline II basins, and leads to vastly overestimated coupling constants. A comparison of these force fields to one another and to the coil library (Panel D) illustrates the small range of coupling constants obtained using LJ parameters for standard force fields. This finding disagrees qualitatively with the predictions made based on coil libraries. We find excellent agreement between calculations based on parameters using the OPLS-AA/L and AMBER force fields, and this is noteworthy given the differences in the parameters.

Excluded volume interactions based on standard force field parameters (OPLS-AA/L, AMBER, and GROMOS) lead to severe restrictions in (ϕ,ψ)-space.

This was inferred from visual inspection of Ramachandran maps (data not shown) and we concluded that LJ parameters from these standard force fields are not well suited for use with the ABSINTH model. This conclusion is justified based on the observations that: i) all coupling constants are too large and ii) there is little to no sensitivity with sidechain type.

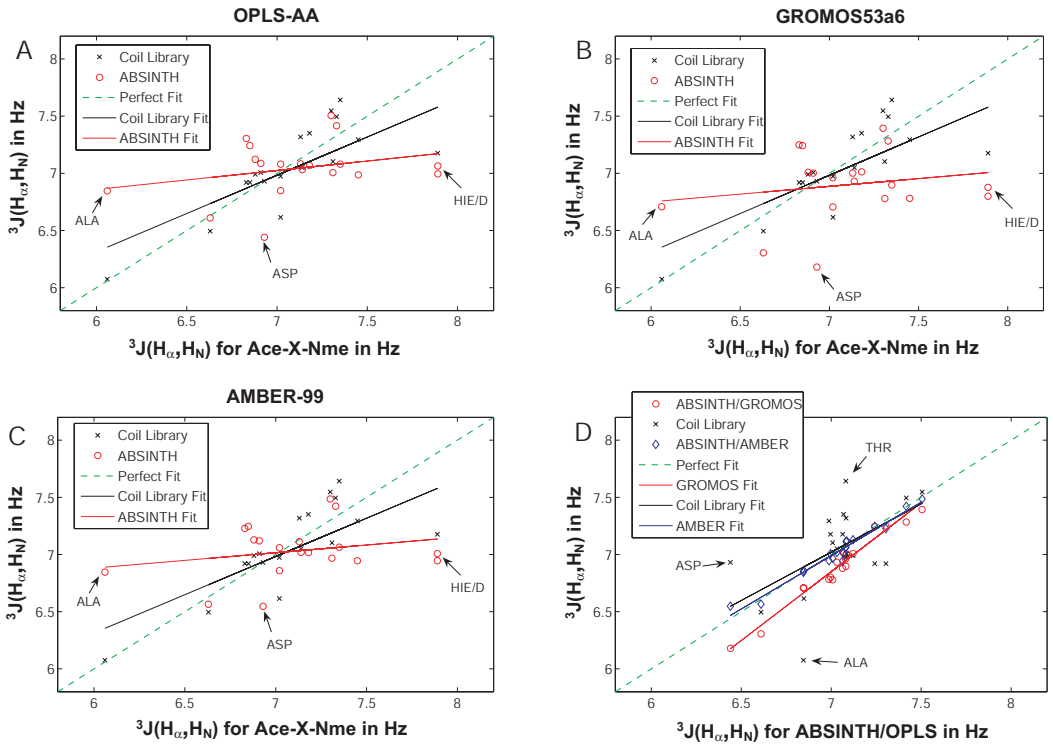

Figure 3.5: **NMR $^3J(H_\alpha N_H)$ coupling constants obtained using ABSINTH's continuum solvation model coupled to modified LJ parameters and standard partial charge sets**. Panels A, B, and C show the comparison of experimental values to the coil library values as well as the simulated results for the OPLS-AA/L (A), the GROMOS (B), and the AMBER (C) charges, respectively. Panel D shows a comparison between the values obtained with OPLS-AA/L to the other computational as well as the coil library data. Drastic outliers are indicated on the plots.

Figure 3.5 shows that we are able to remedy the deviation between the different parameter sets, irrespective of charge set used, by using a consistent LJ parameter set which is detailed in Table 3.b. These parameters are based on atomic radii in small molecule crystals[53] and on generic choices for the interaction strengths, intended to mimic values used in standard force fields.[54,55,63] As is apparent, these parameters coupled to any of the three charge sets (Panels A, B, and C) provide better agreement and a larger sensitivity with respect to residue type. Prominent outliers with respect to the experimental values are alanine, aspartic acid, and histidine. Similarly, outliers with respect to the coil library are alanine, threonine, and aspartic acid, which are indicated in Panel D. Panel D also shows that we observe extremely good agreement for coupling constant values using charge sets from independent force fields.

Within the continuum solvation model adopted in ABSINTH, steric interactions dominate the preferences for the ϕ-angle. Therefore, we are able to remedy deviations in local steric preferences by using a different, consistent set of LJ parameters with all three charge-sets and the hallmark of these LJ parameters are the smaller values for hard sphere radii. The only consistent and drastic outlier is alanine for which we currently have no convincing explanation. The extremely low coupling constant seen experimentally suggests dominant population of the polyproline II- and α-basins, much more so than for any other residue type. Such a strong preference is inconsistent with the broadness of distributions in ϕ/ψ-space we generally observe in our simulations. Most other outliers involve charged residues for which there typically is more variation in

experiments as well, such as a significant dependence on pH,[84] which is difficult to represent in our continuum model. We also simulated capped pentapeptides with the sequence construct $(Gly)_2$-Xaa-$(Gly)_2$ for which there are experimental data under denaturing conditions.[85] Coupling constants are known to be insensitive to the presence of denaturant;[85] hence we simulated these pentapeptides using the ABSINTH continuum solvation model. The calculated coupling constants obtained for residue Xaa in the context of flanking glycine residues are similar to those obtained for dipeptides (data not shown). This corroborates our conclusion that the LJ parameters are crucial for determining short-range structural preferences, which contribute to the measured values for vicinal coupling constants.

One could argue that the above result is due to the general absence of torsional parameters in ABSINTH, although there are exceptions as described in III.3.6. These parameters describe barriers and staggered conformations for rotations about bonds within polypeptides and one might question the validity of their omission. It has been noted that improvements in torsional parameters are crucial for quantitatively accurate descriptions of conformational equilibria for polypeptides.[65] To test our approach, we calculated conformational populations for alanine dipeptide and compared our results to those obtained by Hu *et al*.[86] These authors analyzed conformational equilibria for glycine and alanine dipeptides using a hybrid quantum mechanics / molecular mechanics (QM/MM) approach. They modeled intra-peptide interactions using the self-consistent charge density functional tight binding (SCC-DFTB) method, whereas peptide-

solvent and solvent-solvent interactions were described using standard molecular mechanics models. They compared their results to those obtained using a range of molecular mechanics force fields with explicit solvent. None of these agreed with the conformational distributions calculated using the QM/MM approach. They also noted that conformational distributions calculated with different molecular mechanics force fields did not agree with each other.

Table 3.e shows conformational populations for alanine dipeptide, calculated using ABSINTH that are compared to those obtained by Hu *et al.* from their QM/MM calculations as well as their molecular mechanics calculations using different force fields:

	QM/MM	**QM/MM**	**AMBER**	**CHA.**	**CEDAR**	**OPLS**	**ABS.**
β	0.48	0.48	0.16	0.50	0.71	0.69	0.50
Pass	0.16	0.14	0.00	0.00	0.00	0.06	0.09
α-R	0.27	0.33	0.84	0.50	0.22	0.25	0.39
α-L	0.07	0.03	0.00	0.00	0.05	0.00	0.01
State 4	0.01	0.01	0.00	0.00	0.02	0.00	0.01
$RMSD_I$	0.00	0.08	0.68	0.29	0.29	0.24	0.15
$MaxD_I$	0.00	0.06	0.57	0.23	0.23	0.21	0.12
$RMSD_{II}$	0.08	0.00	0.62	0.22	0.29	0.24	0.08
$MaxD_{II}$	0.06	0.00	0.51	0.17	0.23	0.21	0.06

Table 3.e: Comparative analysis of conformational statistics for alanine dipeptide. Data for conformational statistics shown in columns 2-6 are taken from Tables I and II in the work of Hu *et al*.[86] They are SCC-DFTB with AMBER, SCC-DFTB with CHARMM 22, AMBER, CHARMM 22 (CHA.), and CEDAR. Values for conformational statistics for

OPLS were computed using molecular dynamics simulations. In these simulations, we used parameters from the OPLS-AA/L force field for the peptide and the TIP3P model for water molecules. The simulations were carried out with a single alanine dipeptide in a cubic box of side 25Å. The Berendsen thermostat (T=298K; coupling constant 0.1ps) and manostat[87] (P=1bar; coupling constant, 1ps) were used to simulate the peptide in water in the isothermal-isobaric ensemble. The SETTLE algorithm was used to constrain bond lengths and bond angles for the water molecules, whereas the LINCS method was used to constrain all bond lengths in the peptide. A time step of 2.0fs was used and the equations of motion were integrated using the leapfrog method as implemented in the GROMACS package. A 10Å spherical cutoff was used for both LJ and electrostatic interactions. Neighbor lists were updated once every five time steps and a reaction field with a bulk dielectric constant of 80 was used as a method to introduce corrections due to long-range electrostatic interactions. Data shown in the table are averages over 40 independent simulations, each of length 30ns. Values for conformational statistics for the ABSINTH model were obtained using MMC simulations. Details of the move sets used are shown in Table 3.c. $RMSD_{I}$ is the root-mean-square deviation between statistics shown in columns 2-8 (for the five conformational states) and the statistics shown in column 2 (SCC-DFTB with AMBER). $MaxD_{I}$ is the unsigned maximal deviation between statistics shown in columns 2-8 and the statistics shown in column 2. Conversely, $RMSD_{II}$ is the root-mean-square deviation between statistics shown in columns 2-8 and the statistics shown in column 3 (SCC-DFTB with CHARMM 22) and $MaxD_{II}$ is the unsigned maximal deviation between statistics shown in columns 2-8 and the statistics shown in column 3.

Hu *et al.* reported two sets of QM/MM data that were consistent with each other (SCC-DFTB with either AMBER or CHARMM 22). The two calculations

differed in the choice of LJ parameters used to describe the peptide for modeling peptide-solvent interactions. The QM/MM calculations did not include any empirical torsional potentials because all intra-peptide interactions were described using quantum mechanics:

The results shown in Table 3.e are very encouraging for our approach. When we compare the statistics for specific conformational intervals, it becomes clear that the results obtained using the ABSINTH force field show the best agreement with the QM/MM data. This point is also emphasized when we compare pairwise root mean square deviations between data obtained using different force fields and those obtained using QM/MM. Hu *et al.* also showed that their QM/MM data (and by extension the ABSINTH data) are in good agreement with statistics obtained from the distributions of ϕ,ψ-angles in the protein data bank.[88]

The good agreement between QM/MM data and ABSINTH is very important because it suggests that the description of backbone conformational equilibria using ABSINTH is reasonable. The energy landscape obtained using QM/MM and ABSINTH for alanine dipeptide is in general flatter than what one obtains with the other force fields. It appears that the combination of LJ parameters and stiff torsional potentials in molecular mechanics force fields makes them too restrictive. This in turn might pose challenges for accurate modeling of conformational heterogeneity in IDPs because of significant pre-organization at the level of an individual residue. Given our interest in IDPs as opposed to structure prediction, we propose that the ABSINTH approach might

be a more reasonable alternative for simulating conformational heterogeneity that is characteristic of IDPs.

III.6.2. Thermal Unfolding of two Small Proteins

The 56-residue B1 domain of streptococcal protein G is stable as an isolated construct and characterized by a well-defined α/β-fold and unusually high thermal stability. Its structure has been determined by NMR[89] and the maximum melting temperature was found to be 87°C at a pH of 5.4.[90] The exact melting temperature is strongly pH- and salt-dependent: the stability is expected to be significantly reduced at neutral pH based on a recent, systematic study on a structure-preserving mutant.[91] The B1 domain has been studied extensively by computational methods as well.[92-94] Its α/β-fold, its initial characterization as a prototypical two-state folder, and its outstanding stability suggest that this domain is a useful test case for testing new models.

Ideally, the reversible folding of the B1 domain would be demonstrated by simulating the system from two different initial conditions (randomized vs. folded) over a wide range of temperatures. However, the entire domain folds on the ms-timescale, which is a regime that remains inaccessible to unbiased simulation techniques. Here, we show results of MC simulations of the thermal unfolding of the B1 domain when starting from the folded structure (PDB: 1GB1). At low simulation temperatures, we expect the fold to remain stable, while at high temperatures, we expect full denaturation. The unfolding transition is known to be cooperative; another feature expected to be prominent in plots of folding

measures against temperature. A study of thermal unfolding allows us to test two aspects of our model: first, we can assess if the folded species is a stable minimum for a given Hamiltonian. Second, we assess if the protein shows a cooperative transition between folded and unfolded states, in accord with experimental observations and irrespective of the measure used to assess conformational stability. The second point is rarely addressed in simulation studies since the primary interest often lies in the folded species. To describe phenomena such as folding, assembly, or disorder, however, it is crucial that the folded state is not over-stabilized.

Figure 3.6 shows three different folding measures for the B1 domain of protein G as a function of temperature, the first two of which are based on the root mean square deviation (RMSD) from the PDB structure after superposition, using different subsets of the protein. The third is the radius of gyration (R_g) of the molecule, a quantity used to describe its overall size, *i.e.*, to monitor chain collapse / swelling. Both RMSD measures probe secondary and tertiary structure simultaneously. The thermal stability of the B1 domain has mostly been studied using differential scanning calorimetry and circular dichroism (CD) measurements which have been shown to agree well in general.[90,91] The overall RMSD hence seems like a good candidate to unite the local and global features measured experimentally. Conversely, the R_g measure can only probe overall size and is shown to illustrate the polymeric behavior for this system.

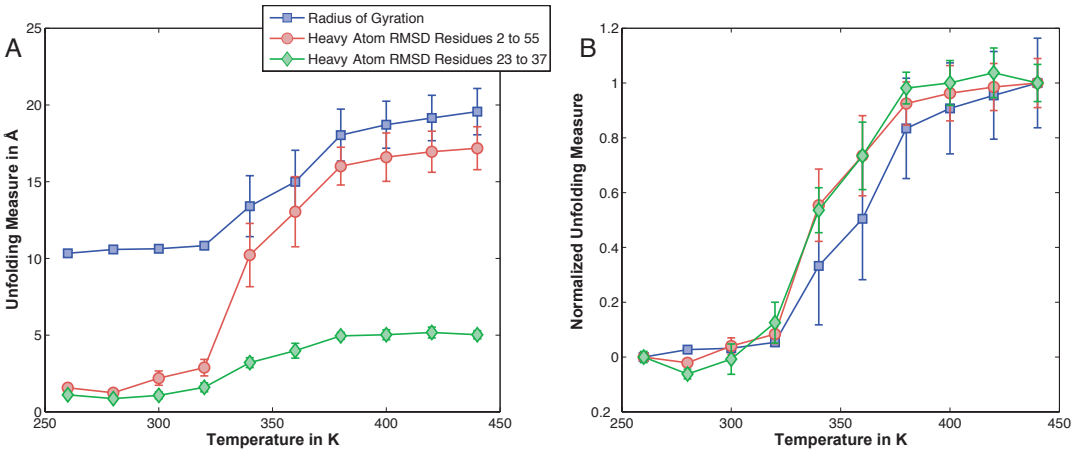

Figure 3.6: **Unfolding measures for the B1 domain of protein G as a function of simulation temperature**. Panel A shows two raw values for the RMSD to the PDB structure: i) for all heavy backbone atoms excluding the terminal residues; and ii) for just the heavy backbone atoms in the helical portion of the protein. It also shows the radius of gyration. The RMSD is based on structural alignments using only the corresponding residues as alignment criteria. Panel B shows values for all three measures normalized to their end points at 260K (0.0, fully folded) as well as 440K (1.0, fully unfolded). Error bars are obtained through block averaging using a block size of 5×10^5 MC steps.

As can be seen in Figure 3.6, all three measures report a cooperative and well-defined unfolding transition with well-defined baselines below 320K and above 380K. Interestingly, when normalized (Panel B), the two RMSD curves coincide almost perfectly indicating that the overall α,β-fold unfolds cooperatively rather than exhibiting disparate stability of the helical and β-sheet parts of the structure. Conversely, the R_g-transition is shifted to slightly higher temperatures indicating that secondary structure melts out partially while the chain remains collapsed. Overall, however, swelling and unfolding are roughly concomitant.

This observation, agrees with the apparent two-state folding behavior reported for this protein.[89,91] More importantly, the melting temperature in the model can be estimated to be around 340K (65-70°C), which is in good agreement with experiment when realizing that the cited 87°C[90,91] are obtained under conditions of maximum stability. The value also coincides well with the number given at low salt and neutral pH for the aforementioned mutant.[91]

To further corroborate that folded proteins are stable minima and show reasonable temperature dependence we chose to study the engrailed homeodomain from Drosophila whose structure was solved to 2.1Å resolution by X-ray crystallography (PDB: 1ENH).[95] It is a three-helix bundle protein which undergoes thermal melting with a midpoint of about 45°C as monitored by CD.[95,96] It serves as a good, complementary test case for the following reasons. First, it is among the fastest folders known to date,[97] which has enabled computer simulations to study the unfolding of this protein directly using MD in explicit solvent on a realistic timescale.[96,97] Second, it has been described as a difficult and hence a good test case for continuum solvation models.[98]

Panel A of Figure 3.7 shows four different unfolding measures which are all based on the RMSD from the PDB structure. If all the proteins heavy backbone atoms are aligned and the RMSD is computed, one obtains a melting curve with a very well-defined upper baseline but a relatively high RMSD of about 3.5Å at low temperature, which continuously grows with increasing temperature. In contrast, if one uses the three helices independently to do the alignment and RMSD computation, helices A and B yield highly cooperative and well-defined

melting transitions with a mid-point of about 330K, while helix C yields a more gradual transition resembling that of the whole protein but shifted to slightly higher temperatures. In Panel B of Figure 3.7 all four unfolding measures are presented in normalized fashion assuming the baseline at low temperature is reasonably flat. As can be seen all measures taken together report on a broad transition region of 300-350K in agreement with experimental data.

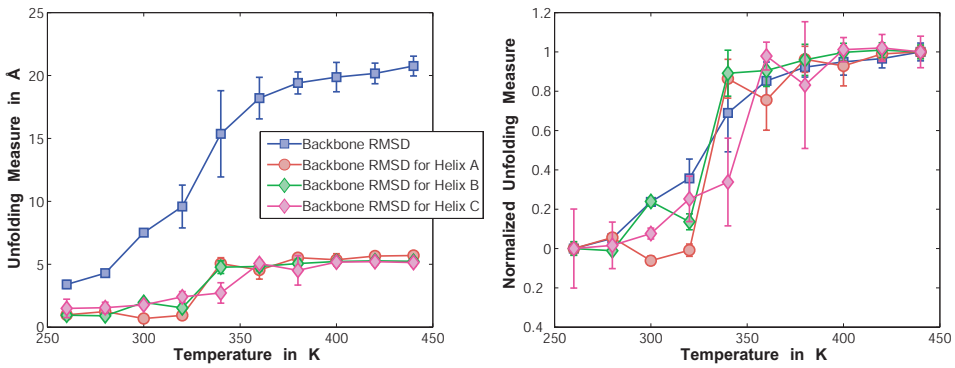

Figure 3.7: **Unfolding measures for the engrailed homeodomain as a function of simulation temperature**. Panel A shows four raw values for the RMSD to the PDB structure which are based on all heavy backbone atoms excluding terminal residues as well as based on the heavy backbone atoms for the three helices individually. Using the PDB-numbering (1ENH), the helices were defined as residues 11 to 25 (A), residues 29 to 41 (B), and as residues 43 to 57 (C). Likewise to Figure 3.6, the RMSD is based on structural alignments using only the corresponding residues as alignment criteria. Panel B shows values for the four measures normalized to their end points at 260K (0.0, fully folded) as well as 440K (1.0, fully unfolded). Error bars are obtained through block averaging using a block size of 5×10^5 MC steps.

We can interpret the data as follows. Unlike for the B1 domain of protein G, the tertiary contacts for the engrailed homeodomain are weak and have substantial residual entropy even at low temperatures. In other words, the relative arrangement of the three helices is not very tightly constrained. With increasing temperatures, tertiary contacts are lost completely but alternative collapsed states with intact helices are visited transiently. This leads to intermediate values for the total RMSD and to large error bars in the transition region. Finally, at high temperature the chain expands fully, and the helices become unstable and melt. This picture obtained from the simulations is consistent with the conclusion from both experiments and computation that the folding of the engrailed homeodomain can be explained using the diffusion-collision model[97,99] in which quasi-stable secondary structure elements "dock" to result in the folded tertiary structure. It is also consistent with the view that the system seems to be much less of a two-state system compared to the B1 domain of protein G and that helix-rich intermediates are populated along the folding/unfolding pathway.[97] Finally, our results somewhat contradict previous simulation work[96] in that we do not find the helices to be significantly populated at high temperatures. It should be noted, however, that the RMSD measure employed here fails to report on small but significant populations of the helical state which we certainly observe for helices A and C, but not for helix B (data not shown), which is in agreement with the literature.[96]

Regarding the reasonable agreement of T_m-values we find with the experimental literature, it must be pointed out that it is well known that simulation

temperatures do not correspond to actual temperatures, because the phase behavior of the solvent is not captured by the continuum. In fact, the proper way to realize temperature dependence in ABSINTH would be to capture the thermal behavior of all the underlying parameters including the reference free energies of solvation (decomposed into entropies and enthalpies), the continuum dielectric, and of course all atomistic parameters. The point here is not to provide quantitative agreement between melting temperatures but to show that the model does not drastically over-stabilize the folded state.

III.6.3. Reversible Folding / Unfolding of the α-Helical FS-Peptide

The 21-residue FS-peptide (Acetyl-A_5[AAARA]$_3$-N-Methylamide) is a member of a class of extremely simple polypeptide systems which undergo a folding transition in aqueous solution. Its melting temperature is estimated to be ca. 305K, *i.e.*, the folded form is expected to be substantially populated at room temperature.[100-102] The α-helical nature of these peptides in the folded form has been established primarily through CD measurements and other spectroscopic techniques.

The FS-peptide is simple and allows us to simulate reversible folding / unfolding transitions as a function of temperature. Additionally, there have been several computational studies on the FS-peptide.[8,78,79] These studies show that the helical form is over-stabilized in simulations with standard force fields, and that *ad hoc* modifications such as the scaling of short-range interactions and the modulation of torsional potentials improve agreement with experimental data.[8,78]

It is worth reiterating that the ABSINTH model does not employ *ad hoc* scaling parameters; nor does it include torsional potentials.

In Figure 3.8, we present the results from 20 independent simulations. For each of the ten temperature values there is both an unfolding simulation starting from the canonical α-helix and a folding simulation starting from a random, extended conformation:

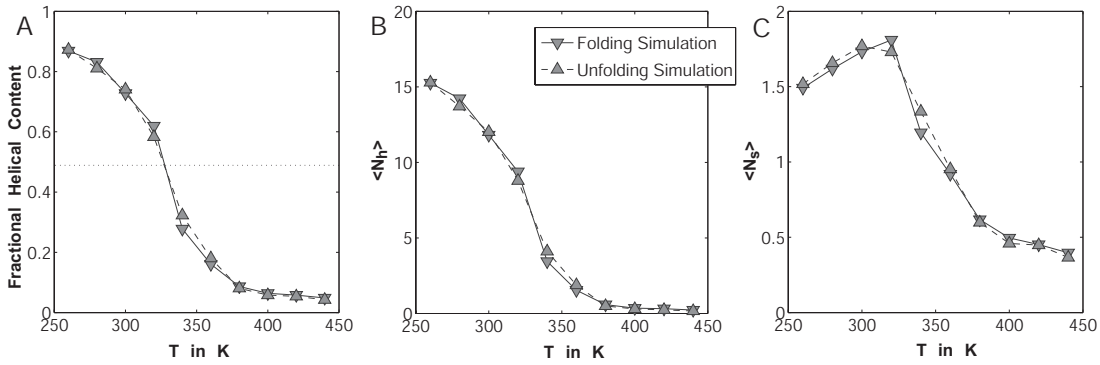

Figure 3.8: **Temperature-induced melting of the FS-peptide**. The temperature-dependence of the fractional helical content (Panel A), the mean number of helical hydrogen bonds (Panel B), and the mean number of helical segments of at least two residues in length (Panel C) are plotted. The folding and unfolding simulations are shown as solid and dashed lines, respectively. The dotted line in Panel A indicates a fractional helicity of 50%, which is used to roughly estimate the melting temperature.

Panel A in Figure 3.8 shows the fractional α-helical content computed according to LR theory. A distinct and cooperative transition is found for both sets of temperature-dependent simulations. We observe virtually no hysteresis between the unfolding and refolding arms indicating that sampling is exhaustive. From the transition region, the melting temperature can be seen to be ~330K,

which is higher than the value of ~305K obtained from experimental studies. However, as Table 3.f shows, the disagreement is smaller vis-à-vis previous computational studies using variants of the AMBER force field with either explicit solvent or a GB/SA continuum solvent description.[78] Panels B and C show how $<N_h>$ and $<N_s>$ vary as a function of temperature (see Equation 3-11). The data indicate that the dominant species at low temperatures is a single, straight α-helix, an observation confirmed by visual inspection of the trajectories (data not shown). The data around 300K are very similar to observations made by Nymeyer and Garcia (see Figures 1 and 2 in their work)[78] using their modified version of AMBER in either explicit or implicit solvent.

Method	T_m **in K**	v **(T in K)**	w
Experiment	~305	0.036 (273)	~1.3
AMBER-94	393 / -	0.27 (300) / 0.36 (305)	2.12 / 1.67
AMBER-GS	342 / -	0.13 (300) / 0.70 (305)	1.67 / 3.70
AMBER-94 / GB/SA	380	0.79 (300)	2.20
AMBER-GS / GB/SA	431	1.57 (300)	4.03
AMBER-99	-	0.06 (305)	0.70
AMBER-99ϕ	-	0.26 (305)	1.26
AMBER-94 -SQ	-	0.28 (305)	1.28
ABSINTH	~330	~0.5 (300)	~1.9

Table 3.f: **Comparative analysis of parameters of the helix-coil transition for the FS-peptide**. AMBER-94 is the full Cornell *et al.* force field[63] while AMBER-GS is the modification introduced by Garcia and Sanbonmatsu.[103] AMBER-99[104] is a more recent version known for disfavoring α-helices while AMBER-99ϕ is the correction introduced

by Sorin and Pande.[79] Finally, AMBER-94–SQ is a further modification introduced by Sorin and Pande shown to illustrate their more extensive study on the impact of non-covalent term scaling.[8] The second column shows the melting temperatures in Kelvin, the third and fourth columns the LR parameters at 300K or 305K, respectively. Two different datasets are shown for AMBER-94 and AMBER-GS which come from Garcia's[78] and Pande's[79] groups, respectively.

Figure 9 shows the LR nucleation and propagation parameters as a function of temperature in Panels A and B. These quantities have been estimated via Equation 3-11 (see III.4) and describe the propensities to populate the α-helical basin in the absence of hydrogen bond stabilization and to extend existing helix nuclei through hydrogen bond-stabilized growth.

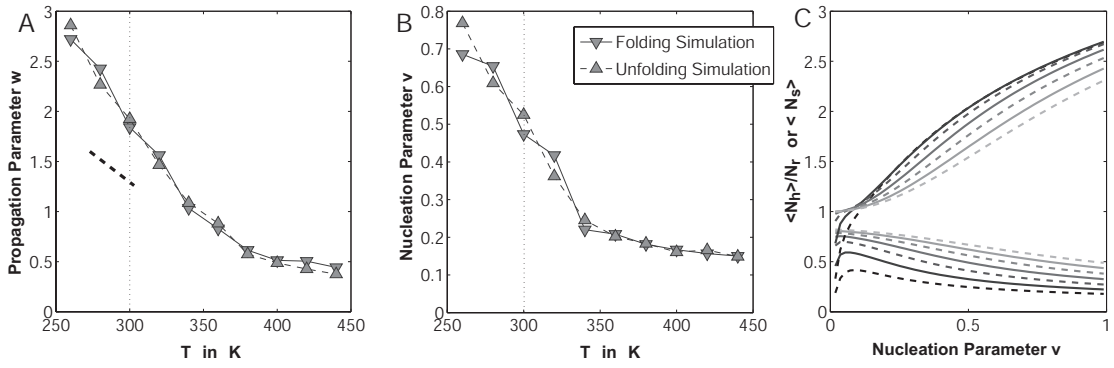

Figure 3.9: **The temperature-dependence of the Lifson-Roig (LR) nucleation** (Panel A) **and propagation** (Panel B) **parameters** shown analogously to Figure 3.8. Dotted lines indicate a temperature of 300K. The thick dashed line in Panel A is the experimentally determined temperature dependence of the propagation parameter. Panel C shows predictions for the mean, fractional number of helical hydrogen bonds (lower set of curves) and for the mean number of helical segments (upper set of curves) from LR theory as a function of the nucleation parameter. A family of curves for values of

w = 1.27, 1.42, 1.57, 1.72, 1.87, 2.02, 2.17 is shown in either case. Increasing values of w are shown as lighter-colored graphs and dashed and solid lines alternate for better clarity.

Panel A shows that w follows the trend for the overall helicity, and that it decreases from values around 2.8 to values around 0.45 throughout the temperature range studied here. Panel B shows that v is temperature-dependent as well, decreasing from values around 0.75 to values around 0.15. Just as we observed for the various measures of helicity in Figure 3.8, the hysteresis between unfolding and refolding arms is minimal indicating very well-converged data. Table 3.f shows that the estimates obtained using ABSINTH are generally comparable to values obtained with other force fields. The work of Nymeyer and Garcia[78] makes it clear that even though v and w might show better agreement with experiment melting temperatures may be overestimated. Furthermore, the experimentally determined temperature-dependence of w[105] is crudely shown as a thick dashed line in Panel A of Figure 3.9. Obviously, ABSINTH slightly overestimates the propagation parameter but it does seem to provide a reasonable representation of the slope. If anything, the latter seems to be slightly overestimated, which stands in contrast to the AMBER-based models deemed most reasonable for which the slope seems to be underestimated.[79] The most relevant comparison in Table 3.f is between the simulations using the GB/SA model and ABSINTH. For this specific system, the latter shows better agreement with experiments than the former, and this holds for all the measures used to quantify helix-coil transitions as shown in Table 3.f.

Two additional points need to be made. As has been noted in the literature, v is generally overestimated by roughly an order of magnitude in simulations. We suggest that this is partly due to the method employed for analyzing simulation data. In experiments, v is obtained through fits to kinetic data on helix nucleation,[106,107] while computationally it is obtained through fits to the equilibrium population of helix segments. Panel C in Figure 3.9 shows predictions from LR theory as a function of v. Clearly, for large enough values of w, high helicity coupled to an average number of helical segments significantly larger than unity (as is usually observed) is only possible if v is substantially larger than the experimentally determined value of 0.036. Even at high temperatures, when entropy dominates, it is impossible to observe very low values for v given the way we compute this parameter from simulation data. This point will be addressed in detail elsewhere. Finally, Table 3.f allows one to make comparisons between the AMBER and OPLS-AA charge sets. It should be noted that the latter were used for the ABSINTH calculations shown here. Gnanakaran and Garcia[9] have observed sharper transitions in their study of a related peptide Ala_{21} using explicit solvent and the OPLS-AA/L force field[65] suggesting that the lack of cooperativity in AMBER might be due to the charge set employed.

III.6.4. The Reversible "Folding" of a β-Hairpin Peptide

For peptides engineered to fold into a β-hairpin, there often is no well-defined transition between the folded and unfolded ensembles. The thermal denaturation of these systems usually shows a broad transition with ill-defined

baselines and little cooperativity.[81,108,109] We can summarize the differences with respect to α-helical peptides as follows:

- The folding of α-helical systems is backbone-driven. This is apparent from the fact that low-complexity sequences such as the FS-peptide do in fact fold and that the simplest chiral residue (alanine) is the one with the largest helix propensity.[110,111] Conversely, the folding of β-hairpins or three-stranded β-sheets is sidechain-driven. This point is made by the fact that design attempts succeed by focusing on optimizing the turn sequence and the sidechain registry.[81,108,112,113] In other words, the chiral peptide backbone of short peptides in aqueous solution shows an intrinsic propensity to populate α-helical but not β-rich conformations. This is illustrated indirectly by the prevalence of ordered, β-rich structures in environments which become less and less aqueous such as protein aggregates[114] or organic solvent mixtures.[115]

- The folding of α-helical systems is well-described by simple models (see above) whereas that of short β-sheet peptides is distinctively heterogeneous and highly sensitive to the experimental probe employed.[109]

- The folded ensemble for α-helical systems is characterized by residual entropy in fraying ends, bending, and possible kinks but always remains well-described by local backbone propensities and the i to $i+4$ hydrogen bond registry.[80] Conversely, the folded ensemble for most β-sheet peptides is almost exclusively constrained by non-local effects such as the arrangement

of sidechains coming from opposite strands. Experimentally, this type of ordering relates to the fluorescence of aromatic residues[81,109] or NMR order parameters such as NOEs.[112,116]

- The kinetics for helix formation are at least an order of magnitude faster than those for hairpin formation.[117] Hence, systems of the latter type pose a much stiffer challenge for computational efforts trying to demonstrate reversible folding.

Most of the simulation studies carried out on β-hairpin peptides have focused on a fragment of the B1 domain of protein G, more precisely the C-terminal hairpin, as it was shown to exhibit features resembling the "native" hairpin experimentally.[116] However, the order parameters chosen in simulation work usually do not relate to experimental probes directly; hence, the relevance of such results for the goal of calibrating force fields is questionable. Moreover, it was recently shown that the NMR data are in fact much more consistent with highly disordered simulation ensembles involving large populations of non-native like structures than with predominantly folded ensembles.[118]

The preceding discussion leads us to choose the so-called tryptophan zippers as our model system. These are very short peptides with two tryptophan pairs on either side which "zip" together to stabilize the β-hairpin conformation.[81] NMR structures could be obtained using distance as well as dihedral restraints. From a simulation standpoint, the system has been studied most extensively using continuum solvation models of the GB/SA flavor.[82,109,119,120] Even in a

continuum solvent, sampling is surprisingly difficult given the small size of these peptides. The system was shown to exhibit heterogeneity both in terms of the kinetics of its thermal unfolding behavior in aqueous solution[109] and in terms of its simulated conformational equilibrium at temperatures, for which experimental data are interpreted to indicate dominant population of the folded basin.[119]

Here, we study "trpzip1" (pdb code: 1LE0)[81] which has the lowest melting temperature among the 12-residue designs but seems to show the cleanest transition between predominantly folded and predominantly unfolded ensembles. Even for this system, however, the experimental data are interpreted to imply that the fraction of folded molecules decreases almost linearly from 0.8 to 0.1 over the wide temperature range of 300 to 360K. Moreover, the maximum folded population is never expected to exceed 0.8, hence indicating substantial residual disorder even in the low-temperature regime. Figure 3.10 shows the temperature dependence of various order parameters for simulations starting from either random extended conformations (folding simulation) or from the NMR structure, more precisely the first model (unfolding simulation). In general, both sets of simulations agree very well with one another. At 300K discrepancies start to arise, and we could not generate hysteresis-free data for temperatures below 300K. This agrees with previous studies which had to use substantially elevated temperatures to achieve converged results for this and similar systems.[119,121]

Panel A of Figure 3.10 shows the mean RMSD, which decreases with decreasing temperature, but only reaches a value of 3.5Å at 300K. Panel B shows the R_g of the hydrophobic cluster driving hairpin formation, *i.e.*, the

sidechains of the four tryptophan residues. This is an order parameter typically used for β-hairpin systems,[122,123] since it addresses the driving force for folding directly. For the tryptophane zippers, however, the NMR structures do not show a true hydrophobic cluster, but instead show the indole rings to be in an edge-to-face arrangement on one face of the hairpin with substantial solvent-accessibility and no stacking or hydrogen bonds. Guvench and Brooks[124] argue that this unusual structure arises due to the electrostatic multipoles in the non-polar parts of the indole rings. The NMR ensemble has a resultant value for the R_g of the hydrophobic cluster of 6.4Å, which is actually larger than what we observe at 300K.

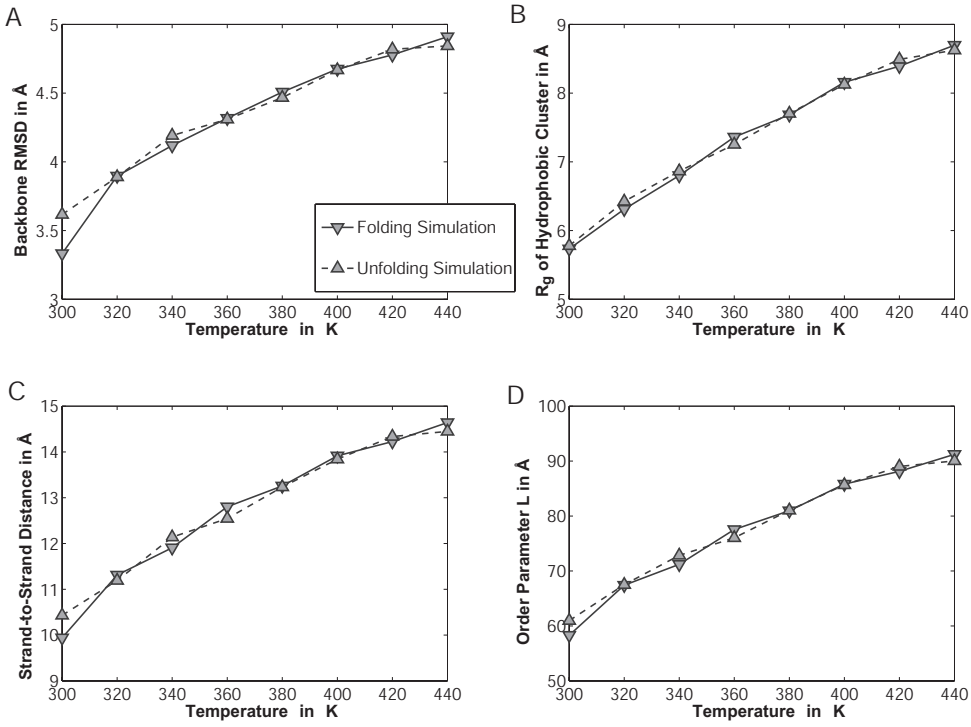

Figure 3.10: **The temperature dependence of various order parameters characterizing the simulated ensembles of the tryptophan zipper "trpzip1"**. Sets of

folding and unfolding simulations are shown as solid and dashed lines, respectively. Panel A shows the heavy backbone atom RMSD to the PDB structure (Model 1 in 1LE0) excluding the N-terminal serine and the C-terminal amide cap. Panel B shows the radius of gyration of the sidechains of the four tryptophan residues. Panel C shows the mean strand-to-strand distance for the perfect hairpin, whereas the order parameter L as defined by Snow *et al.*[82] is shown in Panel D.

Panel C of Figure 3.10 shows the average strand-to-strand distance. The behavior is similar to that seen for the backbone RMSD in that the value for the NMR ensemble (4.8Å) is approached with decreasing temperature but that even at 300K the deviation is quite substantial. Similarly, the order parameter L as defined by Snow and co-workers[82] takes into account native hydrogen bond distances as well as sidechain-sidechain distances for the trypophan pairs found in contact experimentally. Panel D shows that L behaves similarly to both the RMSD and the mean strand-to-strand distance with the NMR ensemble yielding an average L of 27.9Å.

In summary, these results indicate that the ABSINTH Hamiltonian predominantly samples disordered conformations which emphasize the driving force for the collapse of the hairpin but fail to populate the specific structure determined by NMR. In an average sense, a broad basin of structures with native-like features becomes more populated with decreasing temperature, which is in accordance with the experimental data on thermal melting but contradicts the folding estimates deduced from such data.[81] Yang *et al.*[109] have shown for "trpzip2" that by various spectroscopic probes multiple melting transitions can be

identified none of which can be interpreted to uniquely report on the loss of the specific NMR structure as the order parameters in Panels A, C, and D of Figure 3.10 do.

In order to show the differences and similarities between our results and those of other simulation studies we computed two-dimensional potentials of mean force (PMFs) in various combinations of order parameters. Figure 3.11 shows plots analogous to Figure 3a in the work of Snow *et al.*[82] for the folding (Panel A) and unfolding simulations (Panel D) at 300K, respectively:

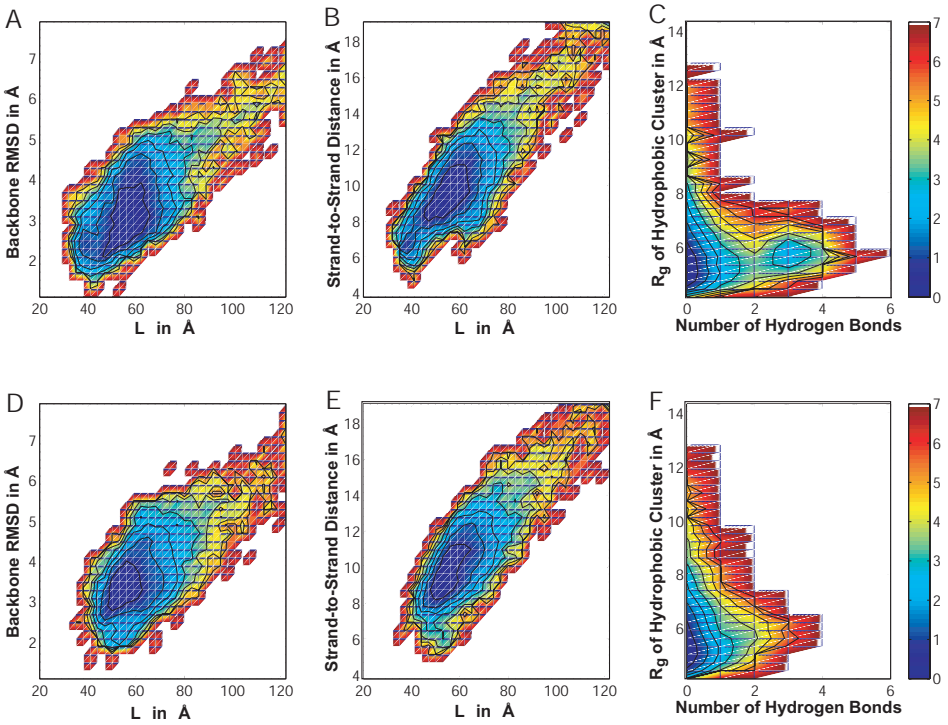

Figure 3.11: **Various two-dimensional potentials of mean force for combinations of order parameters for "trpzip1"** (see III.4 and Figure 3.10). The data are obtained at 300K and are shown for the folding simulation in Panels A, B, and C, and for the unfolding simulation in Panels D, E, and F.

In these PMFs of order parameter L *vs*. the backbone RMSD, the native state would be located in the lower left corner. Clearly, the precise NMR structure is not a relevant part of the free energy landscape. Instead, structures with native-like low L values or low backbone RMSDs are observed independently of one another. This means for example that misfolded hairpins with non-native tryptophan arrangements are observed. In Panel B, we can identify such a misfolded hairpin basin for low values of both the strand-to-strand distance and L, which was not observed to the same extent in the unfolding simulation (Panel E). Panels C and F show PMFs as a function of the number of strand-to-strand hydrogen bonds *vs*. R_g of the hydrophobic cluster and illustrate this point more clearly. For the folding simulation we found a weak, but distinct basin of conformations with substantial hydrogen bonding. In both cases, however, the vast majority of conformations have little to no strand-to-strand hydrogen bonds. It is crucial to point out that in the work of Snow *et al.* the PMFs are created by analysis of a vast number of independent simulations starting from extended states. While they discard a sufficiently long equilibration phase (100ns, which is several times the collapse time), the free energy surfaces are not equilibrated, and the unfolded state is overrepresented.

This leads to the following major conclusions for “trpzip1”:

- The native basin as defined by a specific NMR structure is not a stable conformation for the ABSINTH Hamiltonian. It is noteworthy that Snow *et al*. show that the OPLS-UA force field coupled to the GB/SA continuum solvent is equally unable to stabilize the native basin, and that – unlike common

practice – no 14-scaling was employed in any of their OPLS-AA/GB results. This is an important modification of the force field because the underlying energy landscape may depend strongly on this choice,[8] in particular in a continuum solvent as discussed in III.3.

- A broad basin of states with native-like topology is populated readily and increasingly so with decreasing temperature. This finding suggests that the ABSINTH model can be used to reliably identify native-like basins albeit in a coarse-grained manner.

At this point, we wish to re-emphasize that ABSINTH is not designed as a structure prediction tool. For the applications of interest, it appears more beneficial to underpredict rather than to overpredict the specific structural preferences of polypeptides. While we are actively invested in understanding what components of our model lead to the observed discrepancies for “trpzip1”, we also wish to point out that this result does not imply a general problem of the model in dealing with β-structures. This assertion is supported by our results for the B1 domain of protein G, for which we observe cooperative unfolding in agreement with experimental data indicating that within the context of the full-length protein the hairpin is not destabilized. Therefore we do not pursue tuning of ABSINTH to generate stable hairpins given that the experimental data suggest that such an approach would be unjustified and that the ensembles for such short peptides are indeed heterogeneous.[118]

III.6.5. Polymeric Behavior of Polyglutamine

We have shown, both experimentally[6] and computationally,[125,126] that homopolypeptides composed predominantly of glutamine exhibit a strong preference for collapsed states in aqueous solution. They are intrinsically disordered and have no marked preference for canonical secondary structures. The latter point is supported by experimental results based on CD and NMR spectroscopy[127-129] as well as high-resolution computational studies, *i.e.*, molecular dynamics (MD) simulations in explicit solvent.[126] The collapsed nature of the ensemble can be established through a polymeric scaling law, *i.e.*, the change in size of the molecules with chain lengths. For chains in a poor solvent, *i.e.*, a solvent in which chain-chain contacts are preferred over chain-solvent contacts, collapsed states are preferred and the radius of gyration should scale with chain length N with a scaling exponent of ~0.33 (see II.2):

$$\left\langle R_g \right\rangle = R_o N^{\nu};\ \text{where}\ \nu = \frac{1}{3} \tag{3-12}$$

Here, $\langle R_g \rangle$ is the ensemble-averaged radius of gyration of an individual polypeptide chain, N is the chain length, ν is the actual scaling exponent, and R_o is a parameter related to the monomer size. By plotting $\langle R_g \rangle$ versus N in a double-logarithmic plot, one can obtain the scaling exponent through linear regression. In previous computational work including that presented in Chapter II,[125,126] we were unable to directly measure the scaling exponent according to Equation 3-12 due to the prohibitive cost of such simulations. Instead, we compared the polymeric behavior of Q_{20} in water to two reference models, and established

through alternate means that these chains form collapsed globules and that water is indeed a poor solvent even for such short glutamine-rich peptides.

Figure 3.12 shows the double-logarithmic plot of $\langle R_g \rangle$ versus N obtained using ABSINTH compared to the two reference states employed in Chapter II (see II.3.1 in particular):[125]

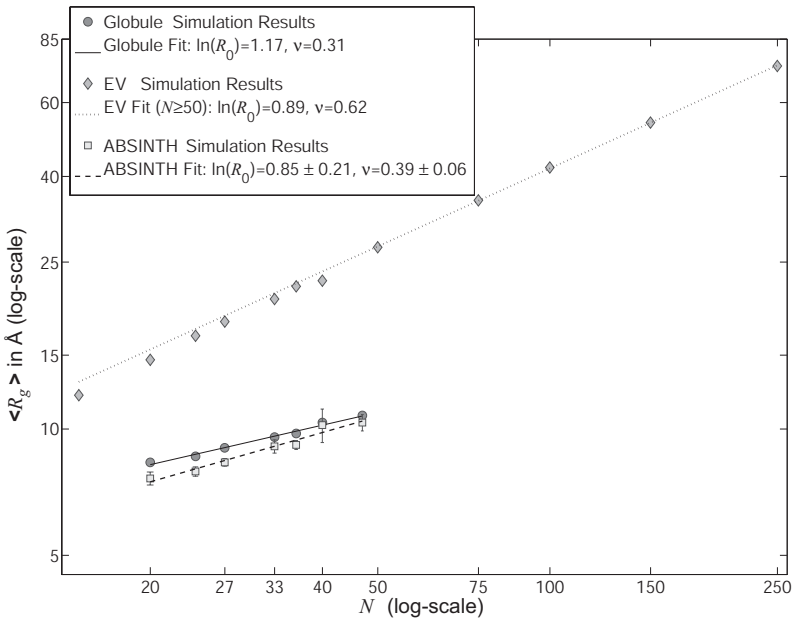

Figure 3.12: **Scaling law for the peptide series Acetyl-(Gln)$_N$-N-Methylamide**. The data obtained with ABSINTH's solvation model is compared to the data for two reference models used in previous work. Error bars on the data for ABSINTH indicate a crude estimate of the reliability of the R_g-values based on the standard deviation of the averages of four independent runs. The uncertainty in the fit parameters was estimated using 50000 independent samples of the data drawn from the estimated normal distributions for each chain length. Due to the crude determination of the parameters for the latter distributions, the numbers are not to be viewed as a rigorous, statistical error estimate.

The preference for collapsed states is preserved in the continuum solvation model and this result agrees with both theory and experiment. Using the uncertainty in the data themselves, we used MC re-sampling of the raw data to obtain an error margin for the scaling exponent of $0.33 \leq \nu \leq 0.45$. Clearly, this is only consistent with poor solvent scaling and not with good solvent scaling, which is observed in the excluded volume (EV) limit as shown in Figure 3.12. Moreover, the experimental results[6] arrive at similar conclusions with regards to the scaling exponent.

However, the scaling exponent is not necessarily the best illustration of solvent quality as small amounts of noise in the data can lead to substantial variability in its estimate. Figure 3.13 shows a more detailed comparison of 30 independent trajectories for Q_{20} to the MD simulations we carried out for the same system (see Chapter II)[125]. We plot the scaling of internal distances (see Equation 2-6) using ABSINTH compared to the calculation in explicit solvent as published. Differences between the two sets of results are mostly statistically insignificant. This suggests that for intrinsically disordered polyglutamine differences in conformational averaging between the implicit and explicit solvent calculations are negligible. Both curves also coincide with the globular reference state indicating that in both explicit and implicit models of solvation water is in fact a poor solvent for these peptides. Furthermore, we analyzed contact maps (data not shown) and concluded that overall there seems to be little to no preference for any kind of consensus secondary structure, even though backbone segment statistics indicate that extended stretches of α-helix are

encountered in a few of the simulations. The observed preference for disorder is in agreement with both experiment and the previous computational studies.

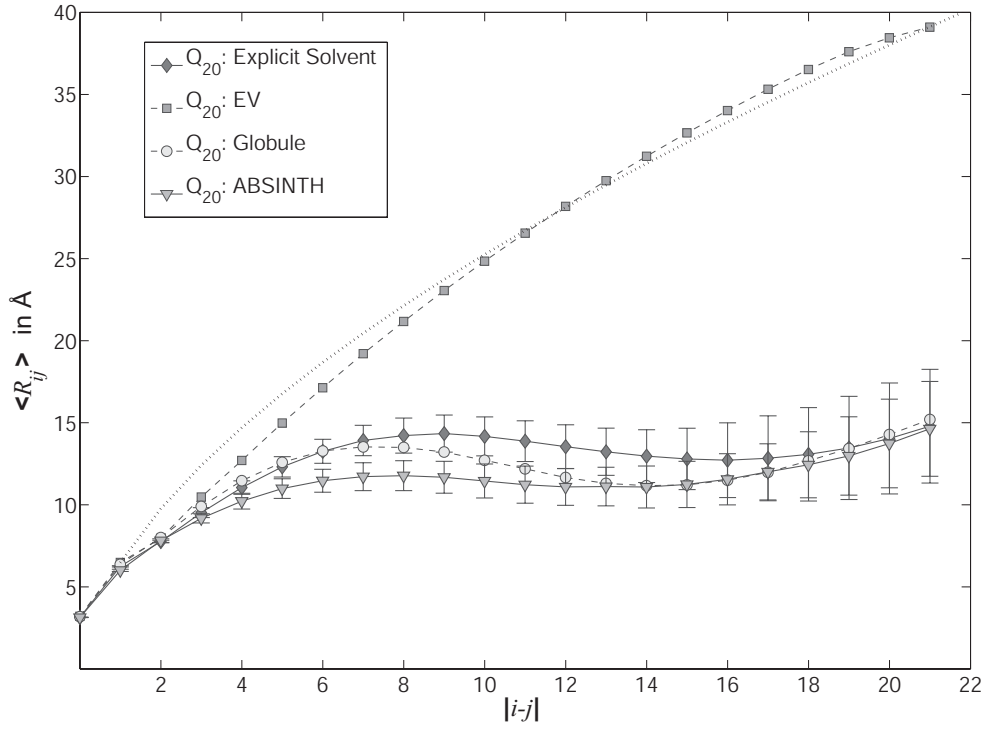

Figure 3.13: **The scaling of internal distances with sequence separation**. The data shown are the results obtained with ABSINTH compared to published results obtained in explicit solvent as well as the two reference models[125]. Error bars are shown for the data in explicit solvent and in ABSINTH and are obtained by calculating the standard deviation of the final averages for each of the 60 and 30 trajectories used, respectively.

III.7. Summary and Conclusions

In this chapter, we have introduced a new continuum solvation model termed ABSINTH. In the broader context of this thesis, a lot of the calibration work performed with the model might appear tangential to the reader. It is very important, however, to point out that i) it is a fundamental necessity for the

relevance of the work presented in subsequent chapters to carry out such detailed tests, and ii) that establishing a model in the broader context of biomolecular applications simultaneously broadens the impact of this thesis for future research which extends far beyond the confines of a particular neurodegenerative disease.

With respect to the second point, the following paragraphs attempt to illustrate why ABSINTH represents a worthy addition to the available continuum solvation models.

ABSINTH is promising

For the test systems analyzed in this manuscript, ABSINTH provides a reasonable description of the underlying physics. Most results are in general agreement with what is known from experiments with two notable exceptions, i) we find specific outliers in the analysis of NMR coupling constants, and ii) we find that the ABSINTH Hamiltonian fails to predict the specific NMR structure for the tryptophan zipper "trpzip1". For the latter, however, the results are not necessarily in fundamental disagreement with the published experimental data as a function of temperature. The test cases here probe the short-range steric preferences of short peptides, the general polymeric nature of Q_{20}, the thermal stability of two small proteins, and reversible folding of both an α-helical and a β-hairpin peptide. Therefore, we conclude that ABSINTH is suitable for simulating processes such as folding/unfolding and self-assembly with semi-quantitative accuracy. The principles underlying phenomena of biological interest are identical; hence, the physical model behind ABSINTH should always apply. We

do have faith that the model can be applied to problems outside of the immediate calibration domain. As an example, other disordered and aggregation-prone systems such as the Aβ-peptide implicated in Alzheimer's disease are currently being investigated in our laboratory using ABSINTH.

As explained in III.3, ABSINTH shares a lot of similarities with the EEF1 model of Lazaridis and Karplus, which has been successfully applied in a variety of contexts[130-133]. ABSINTH, however, does have novel aspects: those include the protocol used to calculate the solvent-accessible volumes; the use of small molecule solutes as solvation groups; the description of partially solvated states; and the screening of Coulomb interactions based on the local solvation environment. The features listed above make ABSINTH a useful model for continuum solvation, which combines aspects of the EEF1 and GB models.

ABSINTH is tunable

It is worth noting that the continuum solvation model can be tuned to change the nature of the solvent. This can be accomplished by varying the solvation parameters r_w, τ_s, χ_s, τ_d, and χ_d. These parameters modulate properties of solvent by tuning the stability of and the cooperativity of transitions between differently solvated states. Similarly, broad changes can be introduced by swapping out parameter sets for the LJ parameters or partial charges as demonstrated in some of our results. It is also possible to carry out simulations including co-solutes such as urea and/or explicit water molecules using the same underlying paradigm, as we have demonstrated for inorganic ions in this work. Finally, there is no fundamental barrier to replace water as the continuum solvent,

as long as the reference free energies of solvation and bulk dielectric are known experimentally for the alternative solvent of interest.[132]

ABSINTH has potential for substantial improvement

All results shown in this manuscript were obtained using MC sampling. Obvious improvements include a switch to a stochastic dynamics approach or even hybrid methods. The treatment of ionic groups as part of the polypeptide and in the bulk provides room for improvement. The goal is to be able to seamlessly integrate the explicit representation of the polymers in aqueous solution with the explicit representations of mobile counterions, which semi-quantitatively capture experimentally observed properties. In addition, the impact of our modified model for short-range electrostatic interactions needs to be analyzed in detail. Possible corrections based on comparison to quantum-chemical data may be required.

Conclusion

We thus conclude that we succeeded in creating a sampling methodology suitable for the questions we wish to ask about the process of polyglutamine aggregation at a physicochemical level. The barrier we encountered and delineated in Chapter II can be breached enough such that our studies extend into chain lengths and system sizes relevant from a disease point of view. Application of the ABSINTH model to this problem is the content of the remaining chapters of this thesis.

III.8. Bibliography

1. Carbone, P.; Varzaneh, H. A. K.; Chen, X.; Müller-Plathe, F. J Chem Phys 2008, 128(6), art. no. 064904.

2. Comon, P. Sig Proc 1994, 36(3), 287-314.

3. Vitalis, A.; Pappu, R. V. J Comput Chem 2009, 30(5), 673-699.

4. van Gunsteren, W. F.; Mark, A. E. Eur J Biochem 1992, 204(3), 947-961.

5. Feig, M.; Brooks, C. L. Curr Opin Struct Biol 2004, 14(2), 217-224.

6. Crick, S. L.; Jayaraman, M.; Frieden, C.; Wetzel, R.; Pappu, R. V. Proc Natl Acad Sci U S A 2006, 103(45), 16764-16769.

7. Moglich, A.; Joder, K.; Kiefhaber, T. Proc Natl Acad Sci U S A 2006, 103(33), 12394-12399.

8. Sorin, E. J.; Pande, V. S. J Comput Chem 2005, 26(7), 682-690.

9. Gnanakaran, S.; Garcia, A. E. Prot Struct Funct Bioinf 2005, 59(4), 773-782.

10. Baker, N. A. Curr Opin Struct Biol 2005, 15(2), 137-143.

11. Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Proc Natl Acad Sci U S A 2001, 98(18), 10037-10041.

12. Simonson, T. Curr Opin Struct Biol 2001, 11(2), 243-252.

13. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. J Am Chem Soc 1990, 112(16), 6127-6129.

14. Onufriev, A.; Case, D. A.; Bashford, D. J Comput Chem 2002, 23(14), 1297-1304.

15. Grycuk, T. J Chem Phys 2003, 119(9), 4817-4826.

16. Ooi, T.; Oobatake, M.; Nemethy, G.; Scheraga, H. A. Proc Natl Acad Sci U S A 1987, 84(10), 3086-3090.

17. Chandler, D. Nature 2005, 437(7059), 640-647.

18. Pierotti, R. A. Chem Rev 1976, 76(6), 717-726.

19. Huang, D. M.; Chandler, D. J Phys Chem B 2002, 106(8), 2047-2053.

20. Wagoner, J. A.; Baker, N. A. Proc Natl Acad Sci U S A 2006, 103(22), 8331-8336.

21. Gallicchio, E.; Kubo, M. M.; Levy, R. M. J Phys Chem B 2000, 104(26), 6271-6285.

22. Gallicchio, E.; Levy, R. M. J Comput Chem 2004, 25(4), 479-499.

23. Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L. J Comput Chem 2004, 25(2), 265-284.

24. Warshel, A.; Papazyan, A. Curr Opin Struct Biol 1998, 8(2), 211-217.

25. Im, W. P.; Lee, M. S.; Brooks, C. L. J Comput Chem 2003, 24(14), 1691-1702.

26. Haberthur, U.; Majeux, N.; Werner, P.; Caflisch, A. J Comput Chem 2003, 24(15), 1936-1949.

27. Ferrara, P.; Apostolakis, J.; Caflisch, A. Prot Struct Funct Gen 2002, 46(1), 24-33.

28. Lazaridis, T.; Karplus, M. Prot Struct Funct Gen 1999, 35(2), 133-152.

29. Privalov, P. L.; Makhatadze, G. I. J Mol Biol 1993, 232(2), 660-679.

30. Mallik, B.; Masunov, A.; Lazaridis, T. J Comput Chem 2002, 23(11), 1090-1099.

31. Patriciu, A.; Chirikjian, G. S.; Pappu, R. V. J Chem Phys 2004, 121(24), 12708-12720.

32. Perchak, D.; Skolnick, J.; Yaris, R. Macromolecules 1985, 18(3), 519-525.

33. Karplus, M.; Kushick, J. N. Macromolecules 1981, 14(2), 325-332.

34. Echenique, P.; Calvo, I.; Alonso, J. L. J Comput Chem 2006, 27(14), 1733-1747.

35. Ben-Naim, A.; Marcus, Y. J Chem Phys 1984, 81(4), 2016-2027.

36. Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K. J Am Chem Soc 2003, 125(31), 9523-9530.

37. Alder, B. J.; Pollock, E. L. Annu Rev Phys Chem 1981, 32, 311-329.

38. Paulaitis, M. E.; Pratt, L. R. In Unfolded Proteins, 2002, p 283-310.

39. Marten, B.; Kim, K.; Cortis, C.; Friesner, R. A.; Murphy, R. B.; Ringnalda, M. N.; Sitkoff, D.; Honig, B. J Phys Chem 1996, 100(28), 11775-11788.

40. Wolfenden, R.; Andersson, L.; Cullis, P. M.; Southgate, C. C. B. Biochemistry 1981, 20(4), 849-855.

41. Wolfenden, R. Biochemistry 1978, 17(1), 201-204.

42. Pliego, J. R.; Riveros, J. M. Phys Chem Chem Phys 2002, 4(9), 1622-1627.

43. Castleman, A. W.; Keesee, R. G. Chem Rev 1986, 86(3), 589-618.

44. Jorgensen, W. L.; Gao, J.; Ravimohan, C. J Phys Chem 1985, 89(16), 3470-3473.

45. Swanson, J. M. J.; Mongan, J.; McCammon, J. A. J Phys Chem B 2005, 109(31), 14769-14772.

46. Lee, M. S.; Olson, M. A. J Phys Chem B 2005, 109(11), 5223-5236.

47. Im, W.; Beglov, D.; Roux, B. Comput Phys Commun 1998, 111(1-3), 59-75.

48. Onufriev, A.; Bashford, D.; Case, D. A. J Phys Chem B 2000, 104(15), 3712-3720.

49. Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. J Phys Chem A 1997, 101(16), 3005-3014.

50. Schaefer, M.; Karplus, M. J Phys Chem 1996, 100(5), 1578-1599.

51. Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. J Phys Chem 1996, 100(51), 19824-19839.

52. Zhou, R. H.; Friesner, R. A.; Ghosh, A.; Rizzo, R. C.; Jorgensen, W. L.; Levy, R. M. J Phys Chem B 2001, 105(42), 10388-10397.

53. Hopfinger, A. J. Conformational Properties of Macromolecules; Academic Press: New York, 1973.

54. Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. J Comput Chem 2004, 25(13), 1656-1676.

55. Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J. J Am Chem Soc 1996, 118(45), 11225-11236.

56. Tran, H. T.; Wang, X. L.; Pappu, R. V. Biochemistry 2005, 44(34), 11369-11380.

57. Villa, A.; Mark, A. E. J Comput Chem 2002, 23(5), 548-553.

58. Chang, J.; Lenhoff, A. M.; Sandler, S. I. J Phys Chem B 2007, 111(8), 2098-2106.

59. Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. J Chem Phys 2003, 119(11), 5740-5761.

60. Shirts, M. R.; Pande, V. S. J Chem Phys 2005, 122(13), art. no. 134508.

61. Udier-Blagovic, M.; De Tirado, P. M.; Pearlman, S. A.; Jorgensen, W. L. J Comput Chem 2004, 25(11), 1322-1332.

62. Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. J Phys Chem B 2007, 111(9), 2242-2254.

63. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. J Am Chem Soc 1995, 117(19), 5179-5197.

64. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. J Phys Chem B 1998, 102(18), 3586-3616.

65. Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. J Phys Chem B 2001, 105(28), 6474-6487.

66. Fitzgerald, J. E.; Jha, A. K.; Sosnick, T. R.; Freed, K. F. Biochemistry 2007, 46(3), 669-682.

67. Vlachy, V. Annu Rev Phys Chem 1999, 50, 145-165.

68. Friedman, H. L. Annu Rev Phys Chem 1981, 32, 179-204.

69. Vitalis, A.; Baker, N. A.; McCammon, J. A. Mol Simul 2004, 30(1), 45-61.

70. Marcus, Y. J Chem Soc Farad Trans 1991, 87(18), 2995-2999.

71. Kang, Y. K.; Nemethy, G.; Scheraga, H. A. J Phys Chem 1987, 91(15), 4105-4109.

72. Engh, R. A.; Huber, R. Acta Crystallogr A 1991, 47, 392-400.

73. Favrin, G.; Irback, A.; Sjunnesson, F. J Chem Phys 2001, 114(18), 8154-8158.

74. Vitalis, A.; Steffen, A.; Lyle, N.; Mao, A.; Pappu, R. V. J Chem Theory Comput 2009, *manuscript in preparation*.

75. Karplus, M. J Chem Phys 1959, 30(1), 11-15.

76. Avbelj, F.; Baldwin, R. L. Proc Natl Acad Sci U S A 2003, 100(10), 5742-5747.

77. Lifson, S.; Roig, A. J Chem Phys 1961, 34(6), 1963-1974.

78. Nymeyer, H.; Garcia, A. E. Proc Natl Acad Sci U S A 2003, 100(24), 13934-13939.

79. Sorin, E. J.; Pande, V. S. Biophys J 2005, 88(4), 2472-2493.

80. Hong, Q.; Schellman, J. A. J Phys Chem 1992, 96(10), 3987-3994.

81. Cochran, A. G.; Skelton, N. J.; Starovasnik, M. A. Proc Natl Acad Sci U S A 2001, 98(10), 5578-5583.

82. Snow, C. D.; Qiu, L. L.; Du, D. G.; Gai, F.; Hagen, S. J.; Pande, V. S. Proc Natl Acad Sci U S A 2004, 101(12), 4077-4082.

83. Case, D. A.; Scheurer, C.; Bruschweiler, R. J Am Chem Soc 2000, 122(42), 10390-10397.

84. Avbelj, F.; Grdadolnik, S. G.; Grdadolnik, J.; Baldwin, R. L. Proc Natl Acad Sci U S A 2006, 103(5), 1272-1277.

85. Plaxco, K. W.; Morton, C. J.; Grimshaw, S. B.; Jones, J. A.; Pitkeathly, M.; Campbell, I. D.; Dobson, C. M. J Biomol NMR 1997, 10(3), 221-230.

86. Hu, H.; Elstner, M.; Hermans, J. Prot Struct Funct Gen 2003, 50(3), 451-463.

87. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. J Chem Phys 1984, 81(8), 3684-3690.

88. Lovell, S. C.; Davis, I. W.; Adrendall, W. B.; de Bakker, P. I. W.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. Prot Struct Funct Gen 2003, 50(3), 437-450.

89. Gronenborn, A. M.; Filpula, D. R.; Essig, N. Z.; Achari, A.; Whitlow, M.; Wingfield, P. T.; Clore, G. M. Science 1991, 253(5020), 657-661.

90. Alexander, P.; Fahnestock, S.; Lee, T.; Orban, J.; Bryan, P. Biochemistry 1992, 31(14), 3597-3603.

91. Lindman, S.; Xue, W. F.; Szczepankiewicz, O.; Bauer, M. C.; Nilsson, H.; Linse, S. Biophys J 2006, 90(8), 2911-2921.

92. Sheinerman, F. B.; Brooks, C. L. Proc Natl Acad Sci U S A 1998, 95(4), 1562-1567.

93. Shimada, J.; Shakhnovich, E. I. Proc Natl Acad Sci U S A 2002, 99(17), 11175-11180.

94. Li, X. F.; Hassan, S. A.; Mehler, E. L. Prot Struct Funct Bioinf 2005, 60(3), 464-484.

95. Clarke, N. D.; Kissinger, C. R.; Desjarlais, J.; Gilliland, G. L.; Pabo, C. O. Protein Sci 1994, 3(10), 1779-1787.

96. Mayor, U.; Johnson, C. M.; Daggett, V.; Fersht, A. R. Proc Natl Acad Sci U S A 2000, 97(25), 13518-13522.

97. Mayor, U.; Guydosh, N. R.; Johnson, C. M.; Grossmann, J. G.; Sato, S.; Jas, G. S.; Freund, S. M. V.; Alonso, D. O. V.; Daggett, V.; Fersht, A. R. Nature 2003, 421(6925), 863-867.

98. Anonymous Reviewer: Personal Communication.

99. Islam, S. A.; Karplus, M.; Weaver, D. L. J Mol Biol 2002, 318(1), 199-215.

100. Thompson, P. A.; Eaton, W. A.; Hofrichter, J. Biochemistry 1997, 36(30), 9200-9210.

101. Lockhart, D. J.; Kim, P. S. Science 1993, 260(5105), 198-202.

102. Ianoul, A.; Mikhonin, A.; Lednev, I. K.; Asher, S. A. J Phys Chem A 2002, 106(14), 3621-3624.

103. Garcia, A. E.; Sanbonmatsu, K. Y. Proc Natl Acad Sci U S A 2002, 99(5), 2782-2787.

104. Wang, J. M.; Cieplak, P.; Kollman, P. A. J Comput Chem 2000, 21(12), 1049-1074.

105. Rohl, C. A.; Baldwin, R. L. Biochemistry 1997, 36(28), 8435-8442.

106. Rohl, C. A.; Chakrabartty, A.; Baldwin, R. L. Protein Sci 1996, 5(12), 2623-2637.

107. Rohl, C. A.; Scholtz, J. M.; York, E. J.; Stewart, J. M.; Baldwin, R. L. Biochemistry 1992, 31(5), 1263-1269.

108. Kortemme, T.; Ramirez-Alvarado, M.; Serrano, L. Science 1998, 281(5374), 253-256.

109. Yang, W. Y.; Pitera, J. W.; Swope, W. C.; Gruebele, M. J Mol Biol 2004, 336(1), 241-251.

110. Rohl, C. A.; Fiori, W.; Baldwin, R. L. Proc Natl Acad Sci U S A 1999, 96(7), 3682-3687.

111. Spek, E. J.; Olson, C. A.; Shi, Z. S.; Kallenbach, N. R. J Am Chem Soc 1999, 121(23), 5571-5572.

112. De Alba, E.; Santoro, J.; Rico, M.; Jimenez, M. A. Protein Sci 1999, 8(4), 854-865.

113. Schenck, H. L.; Gellman, S. H. J Am Chem Soc 1998, 120(19), 4869-4870.

114. Kajava, A. V.; Squire, J. M.; Parry, D. A. D. In Fibrous Proteins: Amyloids, Prions And Beta Proteins, 2006, p 1-15.

115. Das, C.; Nayak, V.; Raghothama, S.; Balaram, P. J Pept Res 2000, 56(5), 307-317.

116. Blanco, F. J.; Rivas, G.; Serrano, L. Nat Struct Biol 1994, 1(9), 584-590.

117. Finkelstein, A. V. Prot Struct Funct Gen 1991, 9(1), 23-27.

118. Weinstock, D. S.; Narayanan, C.; Felts, A. K.; Andrec, M.; Levy, R. M.; Wu, K. P.; Baum, J. J Am Chem Soc 2007, 129(16), 4858-4859.

119. Ulmschneider, J. P.; Ulmschneider, M. B.; Di Nola, A. J Phys Chem B 2006, 110(33), 16733-16742.

120. Chen, J. H.; Im, W. P.; Brooks, C. L. J Am Chem Soc 2006, 128(11), 3728-3736.

121. Pitera, J. W.; Haque, I.; Swope, W. C. J Chem Phys 2006, 124(14), art. no. 141102.

122. Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. Prot Struct Funct Bioinf 2004, 56(2), 310-321.

123. Zhou, R. H.; Berne, B. J. Proc Natl Acad Sci U S A 2002, 99(20), 12777-12782.

124. Guvench, O.; Brooks, C. L. J Am Chem Soc 2005, 127(13), 4668-4674.

125. Vitalis, A.; Wang, X.; Pappu, R. V. Biophys J 2007, 93(6), 1923-1937.

126. Wang, X. L.; Vitalis, A.; Wyczalkowski, M. A.; Pappu, R. V. Prot Struct Funct Bioinf 2006, 63(2), 297-311.

127. Bennett, M. J.; Huey-Tubman, K. E.; Herr, A. B.; West, A. P.; Ross, S. A.; Bjorkman, P. J. Proc Natl Acad Sci U S A 2002, 99(18), 11634-11639.

128. Masino, L.; Kelly, G.; Leonard, K.; Trottier, Y.; Pastore, A. FEBS Lett 2002, 513(2-3), 267-272.

129. Chen, S. M.; Ferrone, F. A.; Wetzel, R. Proc Natl Acad Sci U S A 2002, 99(18), 11884-11889.

130. Steinbach, P. J. Prot Struct Funct Bioinf 2004, 57(4), 665-677.

131. Ma, B.; Nussinov, R. Protein Eng 2003, 16(8), 561-575.

132. Lazaridis, T. Prot Struct Funct Gen 2003, 52(2), 176-192.

133. Lazaridis, T.; Mallik, B.; Chen, Y. J Phys Chem B 2005, 109(31), 15098-15106.

CHAPTER IV. THE EFFECTS OF CHAIN LENGTH AND SOLVENT QUALITY ON CONFORMATIONAL EQUILIBRIA AND DIMERIZATION OF POLYGLUTAMINE

IV.1. Preamble

Progress in science is rarely linear. As might be concluded from the magnitude of Chapter III, the calibration process for the ABSINTH model took an extraordinary amount of time and effort. However, we did not want to lose sight of our biological interests. Hence, work on the dimerization of polyglutamine peptides began long before all the work pertaining to Chapter III was completed. Xiaoling Wang, at that time a postdoctoral researcher in the laboratory, had already been successful in establishing and testing the simulation protocol followed in Chapter II. Her contributions to the work in this chapter are similar: She carried out all of the groundbreaking work needed to establish a reliable protocol which would allow the equilibrium sampling of the association of two polyglutamine peptides across a range of chain lengths relevant in the disease context (Q_5, Q_{15}, Q_{30}, and Q_{45}) and under a range of solvent qualities using temperature as a universal control. She performed preliminary analyses and – unlike in Chapter II – here she also set up and ran the simulations we ultimately analyzed for the manuscript to emerge from this work.[1]

Given this "historical" context, it is not overly surprising that the data presented in this chapter carry two minor caveats. First, as detailed in IV.3.3, we

employed a slightly modified version of the ABSINTH force field as introduced in Chapter III.[2] This was simply a result of the calibration efforts not being finished when the work presented here started. Second, as is also detailed in IV.3, we used a sampling protocol not guaranteed to quantitatively preserve sampling of the canonical ensemble at all temperatures. By allowing the replica-exchange (REX) protocol to swap structures between all solvent qualities, a minor bias is introduced due to asymmetric mixing. This bias error leads to perturbations, which lets transitions along the exchange parameter (temperature) appear slightly broader than they might actually be.

The criticism as to why we chose not to use a more rigorous sampling protocol is answered as follows: the simulations presented here are of substantial cost and cannot easily be repeated. This obvious resource limitation motivated us to find a compromise between rigor, *i.e.*, a minimization of any bias errors, and accuracy, *i.e.*, a minimization of any statistical errors. Due to the inherent complexity of sampling conformational equilibria and association of peptides of up to 45 residues in length, we willingly traded a small bias error for a significant gain in statistical accuracy. This is justified twofold: first, all results presented in this chapter are qualitative in nature and minor quantitative deviations do not affect its conclusions in any way. Second, more recent work[3] has shown that the impacts of both caveats stated in the preceding discussion are in fact minor. This work is presented in Chapters V and VI.

IV.2. Introduction to the Association of Homopolymers

The pathogenic features of CAG repeat diseases are typically interpreted to be predominantly triggered by the polyglutamine stretches found in the expanded disease proteins (see I.2). This is due to the qualitatively similar phenotypes exhibited by diseases affecting completely unrelated proteins. We can therefore formulate a model system *in silico* in which we address the *intrinsic* properties of peptides composed exclusively of glutamine. By stripping away native flanking sequences, the problem is reduced to that of studying the conformational equilibria and aggregation of a homopolymer in a poor solvent.

The connection between the phenomenology of polyglutamine aggregation and the well-established field of conformational and phase equilibria of synthetic homopolymers has been made in a recent review article.[4] This line of thought originates in the seminal works of Flory,[5,6] Huggins,[7,8] and others.[9-12] In a poor solvent, polymers form homogeneously mixed solutions of isolated globules under dilute solution conditions. As concentration increases, the system enters the two-phase regime where there is a clear driving force for phase separation, *i.e.*, aggregation. The poorness of the solvent is now exemplified in its expulsion from a polymer-rich phase. Conversely, in the single molecule limit, intra-chain 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 (see II.2 and III.6.5).[13,14] In poor solvents, the stabilities of collapsed structures and the spontaneities of homotypic intermolecular associations – both of which are governed by attractive

two-body interactions – will increase with chain length.[4] We directly test this last statement for a particular homopolymer – polyglutamine – in this chapter.

Departure from such general homopolymeric behavior may be expected if the polymer is prone to the formation of specific structures in a chain-length dependent fashion. This occurs much more readily for block copolymers. For example, the self-assembly of spider silk block copolymers[15] or collagen microfibrils[16] would be ill-described by the basic tenets outlined above. Monomeric polyglutamine constructs, however, are intrinsically disordered and this holds true irrespective of chain length.[17,18] As would be expected from a simple model, the spontaneity and overall rate of aggregation increase systematically with chain length.[19,20] Under certain conditions, synthetic peptides rich in glutamine form large aggregates with many morphological and dye-binding characteristics that mark these aggregates as being amyloid-like.[21,22] This latter point is important since it suggests that there are structural signatures associated with the aggregation of polyglutamine – at least under certain solution conditions. Work which analyzes the impact of the characteristic β-secondary structure on the thermodynamics of polyglutamine dimerization is presented in Chapter V.

Here, we seek to learn more about the mechanisms by which polyglutamine molecules self-associate to form aggregates.[19] Computational work[23] presented in Chapter II and results from FCS studies[24] helped establish that aqueous milieus at ca. 25°C are poor solvents for polyglutamine. This defines a simple and generic driving force for aggregation. Is it then reasonable

to assume that polyglutamine aggregation follows homogeneous nucleation as suggested?[19] Is the formation of a specific, high energy, conformational species at the monomer level a pre-requisite for aggregation?[25] The number of chain molecules within an aggregate can vary and the smallest "aggregate" is a dimer. If the mechanism of aggregation strictly follows the above model and we find dimerization to be spontaneous, then there are two possibilities: i) the monomeric form of polyglutamine must be the critical nucleus; or ii) the observed dimer must be an off-pathway event. Scenario ii) is much more likely but also gives rise to an alternative explanation: theoretical work suggests that polymer aggregation in poor solvents does not follow homogeneous nucleation.[10] Dimer formation might instead be indicative of heterogeneous pathways leading to a wide distribution of soluble oligomers. The latter may well represent either on-pathway intermediates or off-pathway states of a much more complex aggregation mechanism.

Here, we wish to begin to distinguish between conflicting suggestions, and hence we interrogate intermolecular associations and conformational preferences realized at low concentrations and low copy numbers via computer simulation. The ABSINTH framework[2] allows us to employ temperature as a smooth dial of solvent quality. The characterization of system properties as a function of solvent quality will allow us to adjudicate how much polyglutamine behaves like a generic homopolymer.

The rest of this chapter covers the following material: first, we introduce the computational methods employed here (IV.3). Next, we present results from quantitative studies of the length dependence of coil-to-globule transitions for

monomeric polyglutamine molecules (IV.4.1). This is followed by quantification of the length and temperature dependence of monomer-dimer equilibria (IV.4.2). We then focus our analysis on the correlation between collapse and dimerization (IV.4.3) as well as the driving forces and conformational requirements for both processes (IV.4.4 - IV.4.7). In the discussion section (IV.5), we summarize our results and place our findings in the context of the existing body of experimental data and proposals for mechanisms of polyglutamine aggregation.

IV.3. Simulation Details

IV.3.1. System Setup

Capped polypeptides composed exclusively of glutamine residues (Acetyl-$(Gln)_N$-N-Methylamide abbreviated as Q_N) were built with fixed bond lengths and angles according to the Engh-Huber high-resolution, crystallographic geometries. Peptides with chain lengths N = 5, 15, 30, and 45 were simulated in the nVT ensemble in a spherical droplet of radius R_{droplet}=200Å, where n denotes the number of individual molecules, V the volume of the simulation droplet, and T the simulation temperature. In all cases, polar interactions were truncated at 14Å and short-range steric and dispersive interactions were truncated at 10Å. The cutoffs are justified by the fact that specific interactions between these peptides are exclusively dipolar.

The simulation system consisted of either a single chain (n = 1) for the monomer simulations or of two chains (n = 2) for the dimer case. The chain

molecules were confined to the simulation volume by applying a stiff harmonic boundary potential restraining the molecules to the simulation droplet:

$$E_{bound} = \sum_{\alpha} \begin{cases} 0 \text{ if } \left| \vec{r}_{\alpha} - \vec{r}_{ori} \right| \leq R_{Droplet} \\ k_{bound} \cdot \left(\left| \vec{r}_{\alpha} - \vec{r}_{ori} \right| - R_{Droplet} \right)^2 \text{ else} \end{cases} \quad (4\text{-}1)$$

Here, the sum runs over all the C_α-atoms for all residues, $\vec{r}_{\alpha}$ is the position vector of those C_α-atoms, $\vec{r}_{ori}$ is the position vector of the center of the droplet, and k_{bound} is the stiffness of the one-sided, harmonic restraint.

It is important to note that for a given simulation fluctuations in n are quenched in the nVT ensemble. Therefore, the monomer simulations mimic the infinite dilution limit despite the fact that the simulation volume is finite. For the dimer case, the effective concentration was ca. 100μM, which is in the concentration range of most *in vitro* experiments. Given our simulation setup, there is no possibility of studying the formation of oligomers larger than a dimer. This does not mean that we predict the absence of such larger species in the concentration range of 100μM. Rather, our simulations focus on quantifying the spontaneity of chain length- and temperature-dependent intermolecular associations at low copy numbers and high effective concentrations. This scenario might be reminiscent of *in vivo* settings, although this claim is purely speculative in nature. Extensions to study the formation of larger oligomers will require improvements in sampling methodologies and these are currently being pursued.

IV.3.2. Conformational Sampling

All the simulations presented in this work were performed using Metropolis Monte Carlo (MMC) sampling of the relevant degrees of freedom which for polyglutamine are the ϕ, ψ, and ω angles of the polypeptide backbone as well as the three sidechain dihedral angles (χ_1, χ_2, χ_3) of the glutamine residue. Additionally, for n=2, we include the sampling of rigid body degrees of freedom, namely translations of centers-of-mass and rotational reorientations of molecules. Details of the move sets employed are summarized in Table 4.a:

Move type	**Settings for simulations of monomeric polyglutamine**	**Settings for simulations with pairs of polyglutamine molecules**
Rigid-body	0%	30% (50%, 10Å, 20°)
Omega (ω)	7% (90%, 5°)	4.9% (90%, 5°)
Sidechain (χ_1, χ_2, χ_3)	30% (4x, 60%, 30°)	21% (4x, 60%, 30°)
Backbone ϕ/ψ	63% (70%, 10°)	44.1% (70%, 10°)

Table 4.a: 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. 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 there were a small number of moves for each simulation which were used as swap attempts

for replica exchange. Details on the individual moves are as follows: 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. For Omega-moves there are N+1 ω-angles for a chain length of N due to the acetyl and N-methylamide capping groups. The two sets of values in parentheses are the fraction of ω-moves, which 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, 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. Lastly, the most important moves are those which perturb both the ϕ- and the ψ-angle of a given residue. The values in parentheses are interpreted the same way as for ω-moves.

It is important to remind the reader that appropriate design of MMC move sets allow us to simultaneously probe multiple, disparate length scales rather efficiently, taking advantage of the low overall density. This situation is unlike molecular dynamics sampling which is quite inefficient for sampling large-scale conformational changes as well as intermolecular associations / dissociations. The latter is hindered by slow diffusion and will require adaptive approaches and indeed MMC sampling may be viewed as a variant of such an adaptive

approach. All simulations were carried out using our CAMPARI software package.[26] An overview of all simulations is given in Table 4.b:

Simulation system	Number of independent replica exchange simulations	Simulation temperature of different replicas (K)	Total number of MMC moves that make up a production run	Number of moves between REX swaps
Monomeric Q_5	4	298, 305, 315, 325, 335, 345, 355, 360, 370, 380, 390, 400, 410	2×10^7	0.5×10^5
Monomeric Q_{15}			10^7	
Monomeric Q_{30}			2×10^7	1×10^5
Monomeric Q_{45}			3×10^7	
Q_5 Dimerization	4	298, 305, 315, 325, 335, 345, 355, 360, 370, 380, 390, 400, 410	2×10^7	1×10^5
Q_{15} Dimerization				
Q_{30} Dimerization			4×10^7	
Q_{45} Dimerization			6×10^7	1.5×10^5
Monomeric Q_{15} at higher temperature	1	430, 450, 470, 490, 510, 530, 550	10^7	0.5×10^5
Monomeric Q_{30} at higher temperature	1	430, 450, 470, 500, 550, 600, 650, 700	2×10^7	10^5
Monomeric Q_{45} at higher temperature	1	420, 430, 450, 470, 500, 550, 600, 650, 700	3×10^7	4×10^5

Table 4.b: **Overview of the magnitude of MMC sampling used for polyglutamine**.

As was shown in previous work, intrinsically disordered polyglutamine systems pose a serious challenge for conformational sampling. To improve the quality of our simulation data, we used thermal replica exchange (REX)[27] which adds an extra Markov chain to the MMC sampling. Details of all the parameters for the REX method are summarized in Table 4.b. We improve the overall

efficiency of REX given the finite number of swaps that are feasible during a simulation as follows: In our simulations, we allowed swaps between all unique pairs of replicas in the range 298K$\leq T \leq$410K because the acceptance of proposed swaps between non-adjacent replicas remained finite and we found that this improved the overall quality of sampling, especially for the lower temperatures. A small bias error may be introduced by such a procedure, but the improved statistical accuracy outweighs this concern (see also IV.1).

IV.3.3. Molecular mechanics force field

All of the data presented here were generated using the ABSINTH continuum solvation model.[2] The model was explained in detail in Chapter III. It suffices to point out that it has been shown to reproduce the polymeric behavior of polyglutamine when compared to both simulations in explicit solvent as well as to experimental data (see III.6.5). As alluded to in IV.1, the results presented in the current work were obtained using minor modifications to the published force field. First, for reasons of computational efficiency, partial charges on net-neutral methyl and methylene groups were omitted. All other partial charges are identical to those reported previously and are based on the OPLS-AA/L force field.[28] Second, we employed slightly modified Lennard-Jones (LJ) parameters compared to the parameters published recently (see Table 3.b). The LJ parameters used in this work are shown in Table 4.c. These two modifications were not necessary. As Chapters V and VI show, all of the conclusions regarding the length and temperature dependencies of polyglutamine conformational equilibria and the spontaneities of homodimerization remain qualitatively robust.

However, T_θ shifts to a value that is lower than 410K when we use the parameters that are shown in Table 3.b (see Figures 5.10 and 6.6).

Atom Type	**Example**	**σ_{ii} in Å**	**ε_{ii} in kcal/mol**
N (sp^2)	Amide N	2.70	*0.200*
O (sp)	Carbonyl O	2.70	0.200
C (sp^2)	Carbonyl C	3.00	0.100
C (sp^3)	Methylene C	3.30	0.100
Non-polar H	Methylene H	2.00	0.025
Polar H	Amide H	2.00	0.025

Table 4.c: **Hard sphere diameter and well depth parameters used for computing LJ interactions**. Parameters for cross-interactions were computed using the geometric mixing rule. With the exception of σ_{ii} parameter for sp^2 hybridized nitrogen atoms (see italics) all other parameters are identical to those shown in Table 3.b.

IV.3.4. Data analysis

Most analysis quantities were computed once every 10^3 to 10^4 steps depending on the total extent of the simulation (see Table 4.b). For each monomer / dimer simulation, we carried out multiple, independent simulations with the REX methodology. Therefore, we used a modified block averaging technique to estimate error bars. In this approach, for each temperature point, the data obtained from a single REX simulation run is treated as a single block. With this approach, we are not confronted with the problem of having to chop our simulation data into *ad hoc* blocks. Independence of blocks for averaging is guaranteed because the starting conformations for all simulations are completely

randomized. However, we do not have the resources to carry out hundreds of independent replica exchange runs. Instead, we typically have data from four independent replica exchange runs. Hence, the error bars that result from "block averaging", which in reality is averaging over completely independent trajectories, are not rigorous estimates of statistical and bias errors in sampling; rather, they act as qualitative indicators of the reproducibility of our results between independent runs with randomized starting conformations.

IV.4. Results

IV.4.1. Length Dependence of Conformational Equilibria for Monomeric Polyglutamine

We performed simulations for monomeric polyglutamine as a function of temperature and chain length. Flexible polymers undergo coil-to-globule transitions that are akin to second order phase transitions and characterized by the existence of a "tri-critical" θ-point.[13,14,29] At the θ-temperature ($T=T_\theta$), $\langle R_g \rangle$ is proportional to $N^{0.5}$. Therefore, plots of $\xi = \frac{\langle R_g \rangle}{\sqrt{N}}$ as a function of temperature for different chain lengths should intersect at $T=T_\theta$. Theory also predicts that for $T>T_\theta$ the ratio ξ increases with increasing N whereas for $T<T_\theta$ this ratio decreases as N increases. These predictions are consistent with the fact that for $T>T_\theta$ chain-solvent interactions are preferred in a so-called good solvent: the coil state is favored, and $\langle R_g \rangle$ scales as $N^{0.59}$. Conversely, for $T<T_\theta$, the chain collapses to minimize contacts with the poor solvent and $\langle R_g \rangle$ scales as $N^{0.33}$.

Finally, as chain lengths increase, the sharpness of coil-to-globule transitions should increase and the width of the transition region should decrease.

In Figure 4.1, we plot the variation of ξ as a function of simulation temperature for Q_5, Q_{15}, Q_{30}, and Q_{45}, respectively:

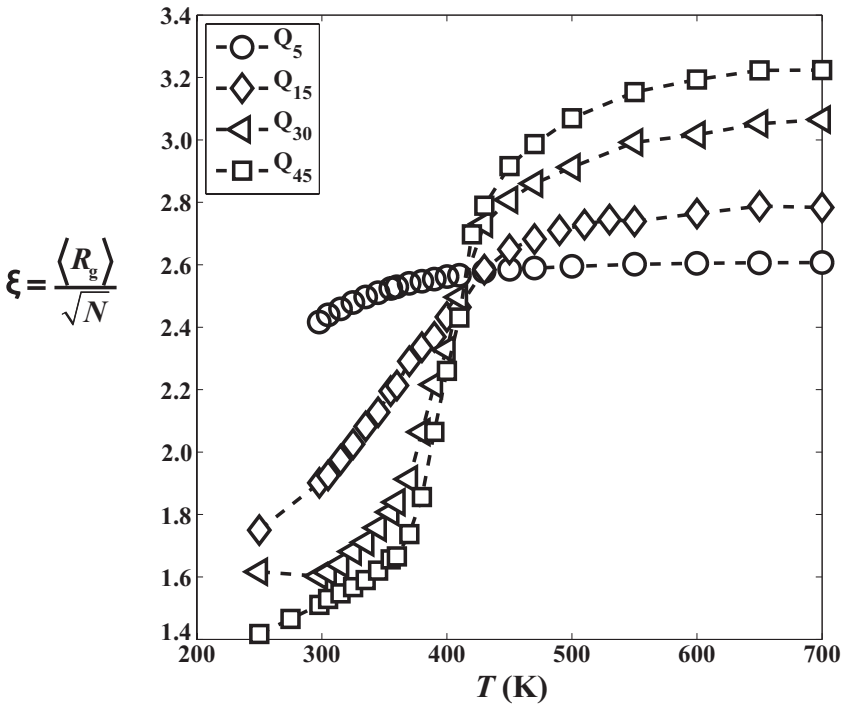

Figure 4.1: **Coil-to-globule transitions for monomeric polyglutamine molecules of different chain lengths** plotted as the variation of normalized chain size ξ (ordinate) with temperature (abscissa).

We find that the coil-to-globule transition is ill-defined for the Q_5 peptide; this is consistent with the concept of "blobs". Within a blob, the balance of chain-chain, chain-solvent, and relevant solvent-solvent interactions is smaller than $k_B T$. Here, T is temperature, and k_B is Boltzmann's constant. If there are k residues in a blob, then the radius of gyration of the blob scales as $k^{1/2}$ and this scaling holds irrespective of solvent quality (temperature). From Figure 4.1 it is clear that Q_5 is

essentially a blob-sized-peptide because ξ does not change significantly with temperature.

From Figure 4.1 we also see that as chain length increases the sharpness of coil-to-globule transitions increases and the width of the transition region decreases. The curves intersect at a common temperature of $T \approx 410$K. For temperatures that lie outside the transition region, 360K$\leq T \leq$430K, ξ decreases with increasing N in the globule limit (T<360K) and it increases with increasing N in the coil limit (T>420K). All of these observations are consistent with expectations listed above from the physics of generic, linear, flexible homopolymers. Such systems collapse in poor solvents in order to sequester themselves from unfavorable interactions with the surrounding milieu. In our calculations, T<360K corresponds to poor solvent conditions.

The results shown in Figure 4.1 suggest that $T \approx 410$K is a reasonable estimate for T_θ. We test this proposal in Figure 4.2 where we plot the scaling of ensemble-averaged internal distances (compare Figures 2.4 and 3.13) as a function of separation in linear sequence. At T_θ, ensemble averages of inter-residue distances $\langle R_{ij} \rangle$ scale as $|j\text{-}i|^{0.5}$. Conversely, for $T < T_\theta$, especially if T is outside the transition region, $\langle R_{ij} \rangle$ for a range of sequence separations should plateau to a constant value. This value is predominantly governed by the density of globules adopted in poor solvents.[29] Ensemble-averaged internal distances $\langle R_{ij} \rangle$ as a function of temperature were calculated as shown below (compare Equation 2-6):

$$\left\langle R_{ij} \right\rangle_T = \left\langle \frac{1}{n_{ij}} \sum_{k \in i} \sum_{l \in j} \left| \mathbf{r}_k^i - \mathbf{r}_l^j \right| \right\rangle_T \tag{4-2}$$

Here, $\mathbf{r}_k^i$ and $\mathbf{r}_l^j$ denote the position vectors of atoms k and l, which are part of residues i and j, respectively; n_{ij} denotes the number of unique pairwise distances between residues i and j and the angular brackets denote an average over all of our simulation data for the system in question at temperature T.

Figure 4.2 shows the variation of $\langle R_{ij} \rangle$ with sequence spacing |j-i| for different chain lengths and temperatures. In Panel A we see that $\langle R_{ij} \rangle$ increases systematically with sequence separation for Q_5 and this holds true irrespective of the simulation temperature. For longer chains and T<360K, $\langle R_{ij} \rangle$ plateaus to fixed values for a range of sequence separations. This temperature regime mimics poor solvent conditions where collapsed states are preferred for monomeric polyglutamine. For T>360K, the data in Panels B, C, and D show that $\langle R_{ij} \rangle$ increases systematically with sequence separation and as T approaches 410K, $\langle R_{ij} \rangle$ scales as $|j\text{-}i|^{0.5}$ with |j-i|. This is demonstrated by favorable comparison of data at T=410K for Q_{15}, Q_{30}, and Q_{45} to dashed curves in Panels B, C, and D that plot $\langle R_{ij} \rangle$ as $R_o|j\text{-}i|^{0.5}$, where R_o=5.7Å is the value of $\langle R_{ij} \rangle$ for |j-i|=1, for all chain lengths and temperatures.

Therefore, for the force field used in this work, T=410K is a reasonable estimate for the θ-temperature (T_θ) for polyglutamine in aqueous solutions. At this temperature, polyglutamine molecules, specifically the longer chains ($N \geq 15$), behave indifferently with regards to their preference for chain-chain versus chain-

solvent interactions. The driving forces for intermolecular associations should be prominent below the θ-temperature. For $T>T_\theta$ or even for temperatures in the immediate vicinity of T_θ there is no *a priori* reason to expect favorable intermolecular associations because chain solvent interactions are favored over chain-chain interactions.

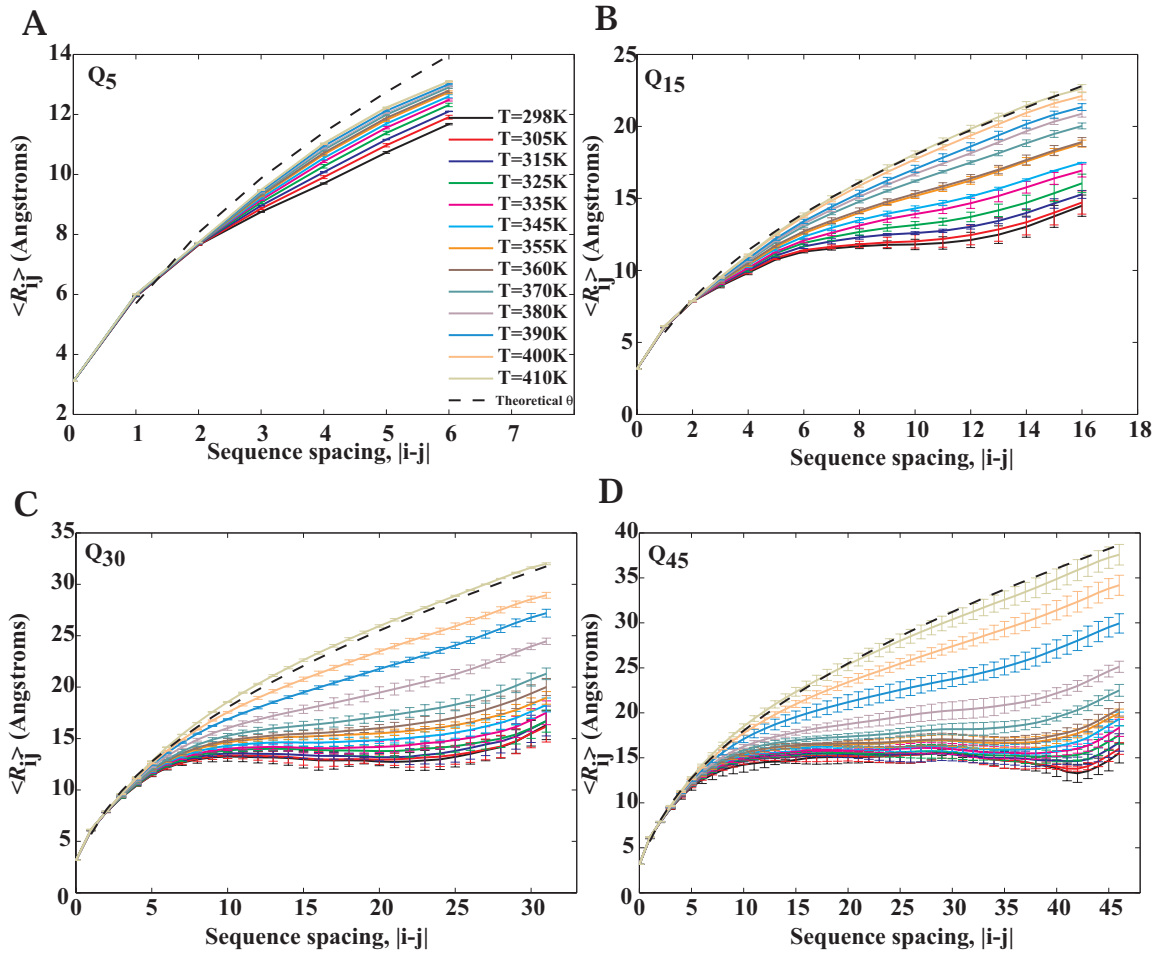

Figure 4.2: **Scaling of average internal distances $\langle R_{ij}\rangle$ between residues i and j as a function of sequence separation**, |j–i|, for monomeric polyglutamine. In each panel, the dashed curve plots $\langle R_{ij}\rangle$ as $R_o|j\text{-}i|^{0.5}$, where R_o=5.7Å, which is the expected profile at T_θ. By comparing data from simulations to the dashed curves we note that T=410K is a reasonable estimate for T_θ for polyglutamine modeled using the ABSINTH force field.

IV.4.2. Length and Temperature Dependence of Spontaneous Homodimerization

Next, we simulated homotypic associations of polyglutamine as a function of chain length and temperature. In addition to the monomer degrees of freedom, *i.e.*, backbone and sidechain torsion angles, rigid body degrees of freedom for each molecule were sampled. Details are presented in IV.3.2.

Figure 4.3 shows temperature dependent cumulative distribution functions $F(R)$ of intermolecular distances for pairs of Q_5, Q_{15}, Q_{30}, and Q_{45} molecules:

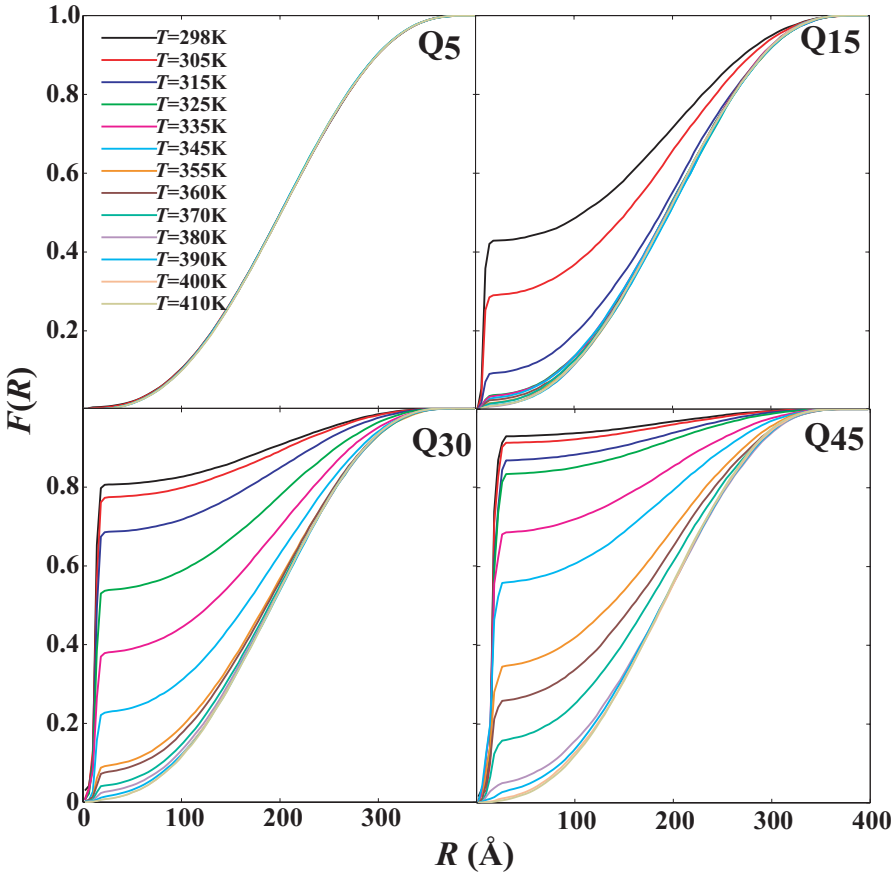

Figure 4.3: **Cumulative distribution functions measuring the probability of sampling specific intermolecular distances between pairs of polyglutamine molecules**.

For a given pair of molecules, $F(R)$ is an estimate of the probability that the average intermolecular separation is less than or equal to R. For Q_5, the cumulative distribution functions are essentially independent of temperature. The likelihood of realizing a specific value of R increases with distance, suggesting that these molecules prefer to diffuse freely about each other. The conclusion is that both conformational equilibria and intermolecular associations for short glutamine-rich peptides are consistent with the behavior of short polar amides in water.

From the analysis of coil-to-globule transitions shown in Figure 4.1 we know that longer chains form more stable globules for T<360K. This chain length dependent drive for intramolecular phase separation has consequences for the spontaneity of intermolecular associations as shown in three of the four panels of Figure 4.3. For temperatures in the range T<360K, the probability of spontaneous homodimerization increases with increasing chain length. For a given value of T, this is quantified in terms of higher probabilities associated with longer chains realizing close intermolecular separations. Conversely, for a given chain length, the probability of spontaneous dimerization decreases with increasing temperature. To quantify these observations, we computed excess interaction coefficients $B_{22}(T)$ using the cumulative distribution functions shown in Figure 4.3. These coefficients are defined as follows:

$$B_{22}\left(T\right)=\frac{\int_{R=0}^{R=D_{\text{droplet}}}\left[F_{T=T_\theta}\left(R\right)-F_T\left(R\right)\right]R^2dR}{\int_{R=0}^{R=D_{\text{droplet}}}F_{T=T_\theta}\left(R\right)R^2dR} \tag{4-3}$$

In Equation 4-3, $F_T(R)$ is the cumulative distribution function at temperature T, $F_{T=T_\theta}(R)$ is the cumulative distribution function at T_θ, and $D_{\text{droplet}}=400\text{Å}$ is the diameter of the droplet used in the simulations (see IV.3.1). The integrals were calculated using an extended trapezoidal rule. The excess interaction coefficients are in the spirit of normalized second virial coefficients that are routinely used in statistical thermodynamics to assess the magnitude of intermolecular associations in solutions of small molecules as well as flexible polymers.[30,31] If $B_{22}(T)$ is less than zero, spontaneous homodimerization is thermodynamically favored vis-à-vis the θ-point and the degree of favorability is assessed by the magnitude of $B_{22}(T)$. If $B_{22}(T)$ is positive, then the chains avoid each other, more so than at the θ-point indicating a clear preference for dissociated states. If the preference for associated and dissociated states is akin to that of an ideal chain, then $B_{22}(T)$ will be zero. It is important to note that $B_{22}(T)$ will plateau to well-defined negative values for molecules which remain associated throughout the entire simulation. It is system size-dependent which does not matter for the context here since $D_{droplet}$ is fixed throughout. Moreover, it is weakly dependent on molecule size given equal associativities with smaller molecules yielding more negative values due to the smaller contact separation.

Figure 4.4 shows two sets of plots. Panel A plots the variation of $B_{22}(T)$ as a function of temperature for $T \leq T_\theta$. Separate curves are shown for each of Q_5, Q_{15}, Q_{30}, and Q_{45}, respectively. For Q_5, $B_{22}(T)$ is negligibly small across the entire temperature range; for the longer chains, $B_{22}(T)$ is negative over different

temperature ranges and the absolute magnitude of $B_{22}(T)$ decreases with increasing temperature. Specifically, $B_{22}(T)$ is negative in the temperature range $T \leq 315$K for Q_{15} and negative in the range $T \leq 360$K for both Q_{30} and Q_{45}. Additionally, at temperatures where $B_{22}(T)$ is negative, its magnitude is greater for longer chains. This is summarized in Panel B which plots the variation of $B_{22}(T)$ as a function of chain length and the different curves denote different temperatures.

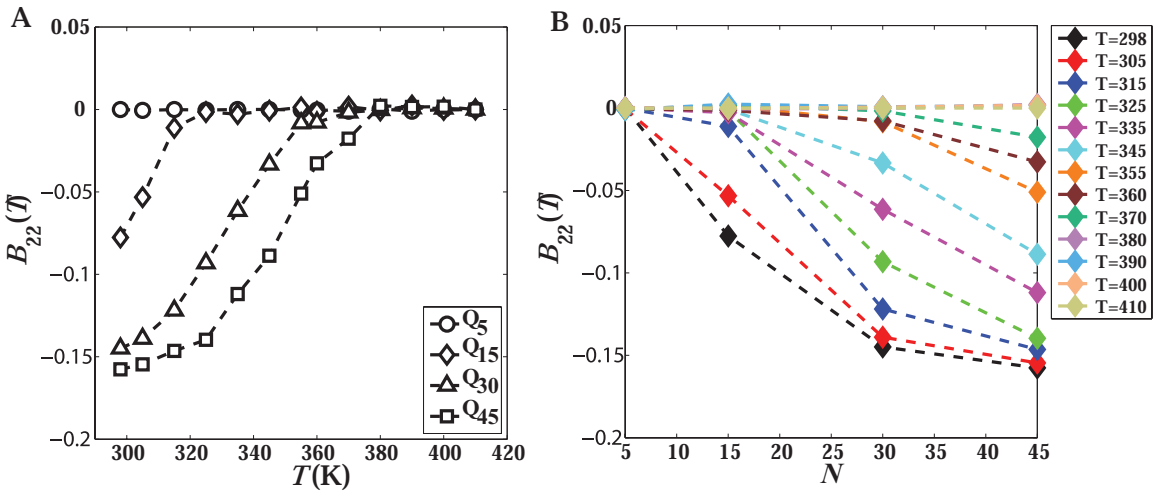

Figure 4.4: **Variation of the intermolecular excess pair interaction coefficients $B_{22}(T)$ as a function of temperature** (Panel A) **and chain length, N** (Panel B).

The results from Figures 4.1, 4.3, and 4.4 may be summarized as follows: the sharpness of coil-to-globule transitions of monomeric polyglutamine increases with chain length. For $T \leq 360$K, Q_{30} and Q_{45} form stable globules. In this temperature range, these peptides also form stable homodimers whose stability decreases steadily with increasing temperature. For a given temperature in the range $T \leq 360$K, homodimers of Q_{45} are more stable than homodimers of

Q_{30}. In contrast, homodimers of Q_{15} are generally less stable and are accessible over a narrower temperature range $T \leq 315$K and this weak dimerization is consistent with the shallow coil-to-globule transition observed for this molecule. The observation of length-dependent dimerization shows that the driving force for polyglutamine aggregation increases with chain length. We now analyze the physical basis for the length and temperature dependence of spontaneous homodimerization in polyglutamine.

IV.4.3. Correlation between Properties of Monomeric Polyglutamine and $B_{22}(T)$

Panels A and B of Figure 4.5 show correlations between the temperature dependencies of specific conformational characteristics of monomeric polyglutamine chains and the temperature dependence of $B_{22}(T)$ in the collapse regime:

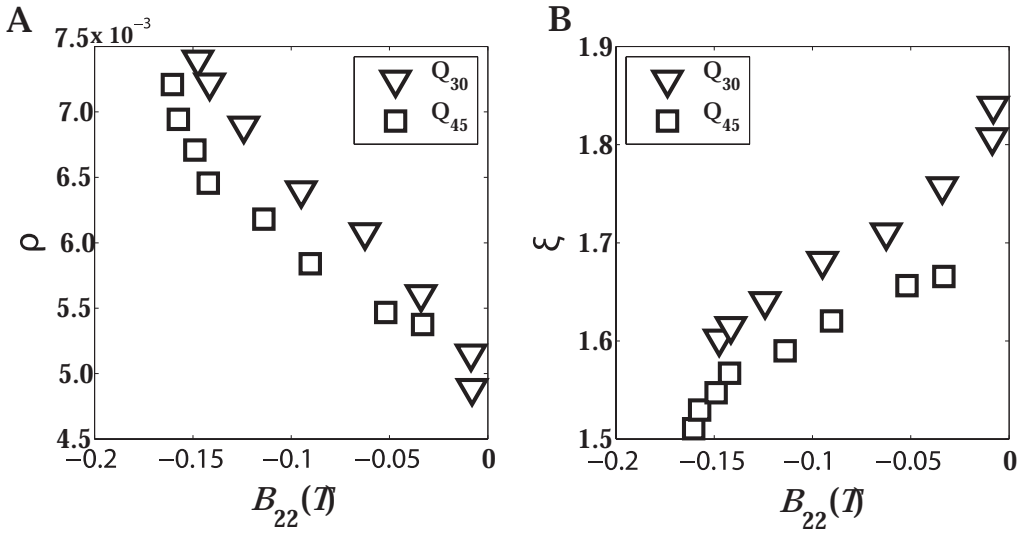

Figure 4.5: **Correlation between monomer properties and $B_{22}(T)$**. Plots showing correlations between temperature dependencies of $B_{22}(T)$ and approximate chain

density, $\rho = \frac{N}{6\langle R_g \rangle^3}$ (units of Å^{-3}) in Panel A and normalized radius of gyration, ξ (units of Å) in Panel B. Data are shown for Q_{30} and Q_{45} in the temperature range 298K$\leq$T$\leq$360K. This temperature regime corresponds to poor solvent conditions for polyglutamine in the ABSINTH model.

Panel A shows a correlation between the density of monomeric globular polyglutamine and $B_{22}(T)$. The magnitude of the latter decreases as density decreases. Similarly, Panel B shows that the magnitude of $B_{22}(T)$ decreases as the ensemble averaged value of R_g increases for monomeric globular polyglutamine.

For a given temperature, the driving forces for chain collapse may be decomposed into two components.[32] Specifically, the mean-field internal energy per residue may be written as:

$$\frac{\langle U \rangle}{N} = C_1(T) + C_2(T) N^{-1/3} \qquad (4\text{-}4)$$

Here, N denotes chain length, $\langle U \rangle$ is the average potential energy at temperature T, $C_1(T)$ measures the bulk energy density, and $C_2(T)$ measures the surface energy density. $C_1(T)$ provides an estimate of the effective strength of self-interactions and $C_2(T)$ estimates the energy associated with making interfaces between polyglutamine and the surrounding solvent. Self-interactions are favorable if $C_1(T)$ is negative and the strengths of favorable interactions are measured by the magnitude of $C_1(T)$. If $C_2(T)$ is positive, its magnitude measures the energy penalty associated with increasing the size of the unfavorable chain-

solvent interface. Conversely, if $C_2(T)$ is negative, then mixing of the chain and solvent is preferred.

Panels A and B in Figure 4.6 plot the temperature dependencies of $C_1(T)$ and $C_2(T)$ in the range $T \leq$360K:

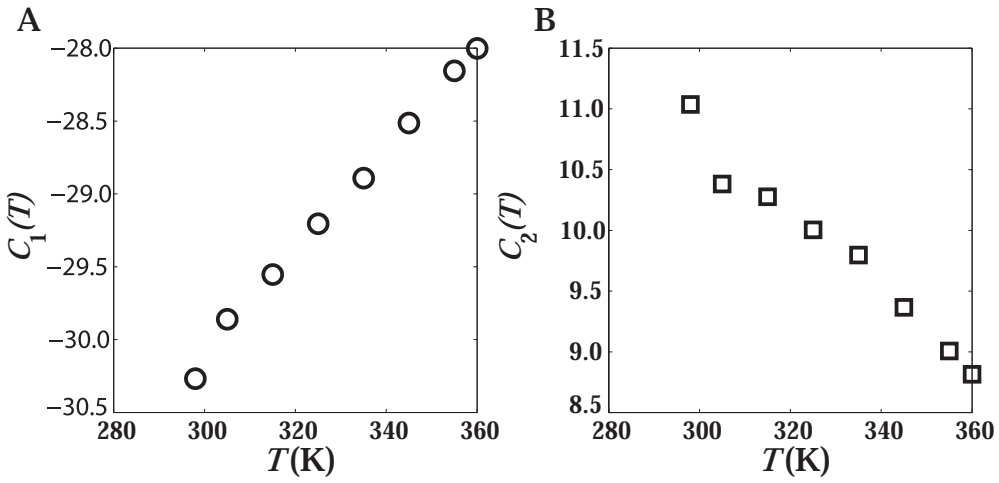

Figure 4.6: **Energy decomposition analysis**. Plots showing the temperature dependencies of the coefficients $C_1(T)$ and $C_2(T)$ that result from linear regression analysis by plotting $\frac{\langle U \rangle}{N}$ vs. $N^{-1/3}$. The coefficient $C_1(T)$ has units of kcal/mol-N, whereas $C_2(T)$ has units of kcal/mol-$N^{2/3}$.

Consistent with previous observations, with decreasing temperature, the stabilities of collapsed states increase. This is indicated by the fact that $C_1(T)$ becomes more negative and $C_2(T)$ becomes more positive. From the correlation analysis in Figures 4.4 and 4.5 and the data in Figure 4.6 we conclude that $B_{22}(T)$ becomes increasingly more negative as the driving force (magnitudes of $C_1(T)$ and $C_2(T)$, respectively) for forming compact (small ξ), dense globules (large ρ) increases. Upon collapse, self-interactions are maximized. Dimerization leads to

a diminution of the unfavorable solute-solvent interface and increased self-interactions through intermolecular association.

Since collapse and dimerization result from the combined drive to minimize unfavorable solute-solvent interfaces and maximize self-interactions, the surface-to-volume ratio (R_{SV}) of a single chain in a poor solvent provides a generic measure of the relevant driving forces.[33] For globules, R_{SV} decreases with increasing chain length because it scales as $N^{-1/3}$. For small N, R_{SV} is large, which means that unfavorable surface energies are not readily offset by favorable self-interactions. As a result, a relatively short peptide like Q_{15} shows weak tendencies toward collapse and stable dimerization. Very few self-interactions may be formed on the inside of a globule for this chain and all residues remain at least partially solvent-exposed. R_{SV} decreases as a function of chain length with $N^{-1/3}$. Hence, it is significantly reduced for Q_{30} and even more so for Q_{45}. From the data we conclude that R_{SV} is small enough to allow for the unfavorable surface energies to be offset comfortably by favorable self-interactions. Visual inspection suggests that these chains form globules with well-defined interiors in which some residues are sequestered from solvent entirely.

Q_{30} and Q_{45} encompass the threshold length range for polyglutamine disease phenotypes.[34] We may speculate that the preceding discussion identifies R_{SV} as a rather simple signature associated with the observed chain-length dependent phenotype: association is phenomenologically coupled to collapse and phase separation may be triggered by a very small nucleus.[19] It is important to point out that such a conjecture is only meaningful because the system is

composed of *highly flexible* polymers. If instead they were rigid spheres that interacted primarily through surface contacts,[35] then the value of R_{SV} for a single chain would be less meaningful because $C_1(T)$ would be irrelevant. In this case, the relevant surface-to-volume ratio would be that of clusters of molecules and not that of a single molecule. The description of polyglutamine aggregation would then follow classical models for homogeneous nucleation where aggregation is favorable only if the cluster size is greater than some critical number.[36]

IV.4.4. Conformational Specificity in Collapse and Intermolecular Associations

The foregoing analysis focused on generic polymer physics parameters and the correlations between these quantities and the spontaneity of intermolecular associations. We also assessed the presence of specific, ensemble-averaged conformational propensities that can be implicated in promoting both collapse and intermolecular associations. Specifically, we asked if there is a discernible increase in β-sheet propensity associated with collapse, spontaneous associations, or both? To answer this question, we computed the fractional α-helical and β-sheet contents using our simulation data.

There are several ways to assess secondary structure content in proteins and polypeptides. We have developed a strategy that is based on analysis of distributions of the ϕ,ψ-angles of the peptide backbone. The resultant measure, shown below, provides a reasonable estimate of secondary structure content as compared to popular measures such as DSSP[37] (compare Figure 5.5). The fractional α and β contents, f_α and f_β, respectively, are defined as:

$$f_{\mathrm{X}} = \frac{1}{N}\sum_{i=1}^{N} f_{\mathrm{X}}^{(\mathrm{i})}, \quad \mathrm{X} \equiv \alpha \text{ or } \beta$$

$$f_{\mathrm{X}}^{(\mathrm{i})} = \begin{cases} 1.0 & \text{if } (\phi_i, \psi_i) \in X \\ \exp\left(-\tau_{\mathrm{X}} d_{\mathrm{X}(i)}^2\right) & \text{otherwise} \end{cases} \qquad (4\text{-}5)$$

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

In Equation 4-5, f_{X} denotes the fractional content of secondary structure type X, where X is either α or β. The $\mathrm{mod}2\pi$ terms corrects for periodicity effects when calculating distances in angular space and N is the number of residues in the sequence, excluding capping groups. The coordinates $(\phi_{\mathrm{X}},\psi_{\mathrm{X}})$ define the reference ϕ,ψ-values to be adopted by an individual residue for the secondary structure motif of type X. If a residue i adopts ϕ,ψ-angles that lie within a circle of radius r_{X}, then the parameter $f_{\mathrm{X}}^{(i)}$ is set to unity; otherwise, $f_{\mathrm{X}}^{(i)}$ assumes a value between 0 and 1, and the precise value is determined by two parameters, *viz*. the distance $d_{\mathrm{X}(i)}$ and τ_{X}. The latter is the width of the Gaussian function used to determine the value to be assigned for $f_{\mathrm{X}}^{(i)}$. For $\mathrm{X}\equiv\alpha$, $(\phi_\alpha,\psi_\alpha)=(-60°,-50°)$, $r_\alpha=30°$, and $\tau_\alpha=0.002\mathrm{deg}^{-2}$. Conversely, for $\mathrm{X}\equiv\beta$, $(\phi_\beta,\psi_\beta)=(-125°,125°)$, $r_\beta=40°$, and $\tau_\beta=0.002\mathrm{deg}^{-2}$.

Figure 4.7 shows four panels that summarize the temperature dependencies of fractional α-helical (f_α) and β-strand (f_β) contents in Q_5, Q_{15}, Q_{30}, and Q_{45}. Data are shown from simulations of monomeric polyglutamine and those with two chains ("dimer"):

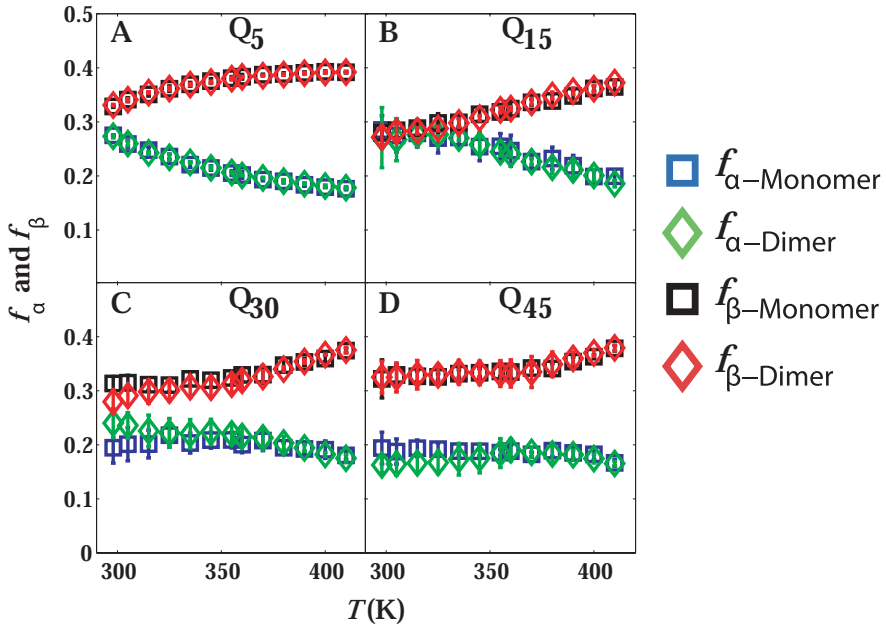

Figure 4.7: **Temperature dependencies of the fractional α-helical (f_α) and β-sheet (f_β) contents for Q_5, Q_{15}, Q_{30}, and Q_{45}, respectively**. The squares (blue and black) quantify f_α and f_β using data from simulations of monomeric polyglutamine. Conversely, the diamonds (green and red) quantify f_α and f_β using data from simulations of pairs of polyglutamine molecules. Data are shown over a range of temperatures spanning the collapse regime and the transition regime up to T_θ.

From the data shown in Figure 4.7 we conclude the following. First, there is a clear, statistically significant diminution in f_α with increasing chain length. Inasmuch as intermolecular associations become favorable with increasing chain length, the decreased α-helical propensities with increasing N suggest a weak correlation between decreased helical propensity and chain associations. Second, while collapse and intermolecular associations show clear temperature dependencies, conformational propensities measured in terms of f_α and f_β show very weak temperature dependencies. Therefore, local conformational

propensities appear to be only weakly coupled to the driving forces for collapse and intermolecular associations. Third, for a given solvent quality (defined by the value of the simulation temperature T), the fractional β-content is greater than or equal to the fractional α-content, and this is true irrespective of chain length. Fourth, of the three chains, Q_{15}, Q_{30}, and Q_{45}, the two longer chains show a diminution in the α-helical content and an enhancement in β-content by a weak increase in f_β with increasing temperature.

Local secondary structure content changes weakly as a function of temperature and chain length. Despite this, the driving forces for collapse and the spontaneities of homodimerization show clear temperature and length dependencies. Therefore, we conclude that disordered globules associate to form disordered dimers. This observation is congruent with the findings of Krishnan and Lindquist[38] for the aggregation of the NM regions of the yeast prion protein Sup35. They found that "molten oligomers", which form as precursors to NM fiber formation, are dominated by contacts between the globular forms of the glutamine- and asparagine-rich N-domain[39] which also forms collapsed structures in its monomeric form.[40]

IV.4.5. Evidence for Intrinsic Disorder in Polyglutamine

Experimental data and computational studies have documented the lack of conformational specificity in monomeric polyglutamine. In our simulation data, this preference for intrinsic disorder prevails despite the preference for collapsed states for temperatures in the range $T \leq 360$K. In Chapter II, we proposed that

intrinsic disorder is a direct consequence of the homopolymeric nature of polyglutamine. The lack of sequence specificity implies that a variety of compact species, irrespective of chain conformation, have equivalent stabilities and that the conformational ensemble therefore is a heterogeneous collection of compact conformations. While this type of disorder is distinct from the disorder associated with denatured proteins, collapse does not imply folding.[41] Our analysis of local conformational propensities makes this point.

Additionally, we can analyze the variations in contact patterns between individual members of the conformational ensemble to assess the degree of disorder as a function of temperature. To accomplish this, we quantify disorder by computing a single figure of merit, namely the normalized variance in the number of intramolecular contacts (σ_N^2) as a function of temperature. This quantity, computed for monomeric polyglutamine, is defined as follows:

$$\sigma_N^2(T) = \frac{1}{NT} \sum_{k=1}^{n_{\max}^{(N)}} \left(k - \langle n_c \rangle_T^{(N)} \right)^2 p_k^{(N)}(T)$$
$$\langle n_c \rangle_T^{(N)} = \sum_{k=1}^{n_{\max}^{(N)}} k p_k^{(N)}(T) \qquad (4\text{-}6)$$

Here, N denotes the chain length, T is the simulation temperature, and $p_k^{(N)}(T)$ is the probability of realizing k intramolecular contacts in a chain of length N at temperature T. A contact is defined by any two non-bonded atoms from residues i and j having a distance less than 3Å. $\langle n_c \rangle_T^{(N)}$ is the average number of

intramolecular contacts in a chain of length N at temperature T and $n_{\max}(N)$ is the maximal number of realizable intramolecular contacts in a chain of length N.

Results for the variation of σ_N^2 as a function of temperature for Q_{15}, Q_{30}, and Q_{45} are shown in Panel A of Figure 4.8:

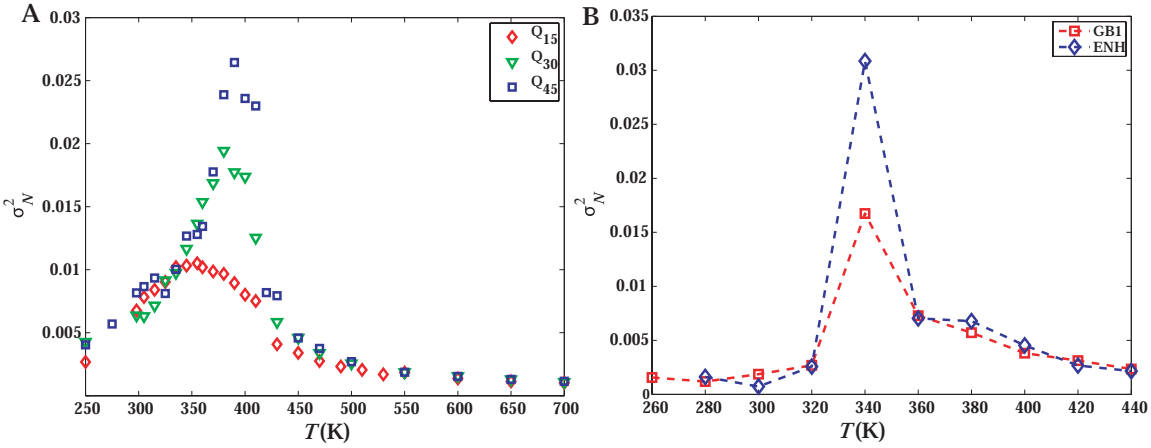

Figure 4.8: **Temperature dependence of the variance in the number of intramolecular contacts** for monomeric polyglutamine of different chain lengths – Panel A – and for two globular proteins with well-defined folds, *viz*. the B1 domain of protein G (1GB1) and engrailed homeodomain (1ENH) – Panel B.

In the high temperature limit (T>450K), past T_θ, chains sample canonical denatured state ensembles where the dominant contacts are local and the likelihood of realizing distal contacts is very small; this is true for a majority of conformations in the ensemble. Consequently, the average number of contacts is small and so is the variance. Just below T_θ, the chains are in the transition region and sample conformations from two distinct ensembles, *viz*. the coil and globule states. In this regime, conformational fluctuations are large and values for σ_N^2 are high because vastly different types of conformations are sampled. In

complete congruence with the analysis shown in Figure 4.1, the sharpness of the coil-to-globule transition increases with N. This is manifest by the fact that as N increases the width of the transition region decreases and the peak height increases in Panel A of Figure 4.8. In the collapse regime, σ_N^2 decreases with decreasing temperature and does not plateau to a well-defined value. This systematic decrease of σ_N^2 with decreasing temperature is a characteristic signature of dynamical disorder and is consistent with the glassy behavior quantified in Chapter II (specifically, see II.4.4).[23] The assignment of dynamical disorder to collapsed polyglutamine is made clear by comparing the variance profiles shown in Panel A of Figure 4.8 to the variance profiles obtained from simulation data for thermal unfolding of two well-folded proteins (shown in Panel B of Figure 4.8), namely the B1 domain of protein G (GB1) and the engrailed homeodomain (ENH) (see Figures 3.6 and 3.7 and III.6.2 as well). There are well-defined baselines in the values of σ_N^2 on either side of the transition region. Additionally, the unfolded baseline (high T) is higher in value than the folded baseline (low T), which is consistent with the adoption of a roughly rigid structure with small-scale fluctuations at low temperature and a heterogeneous ensemble characterized by dominant local contacts at high temperature. In contrast, for polyglutamine, there are temperatures well into the collapsed regime (T<360K) for which σ_N^2 is actually higher than the asymptotic value achieved in the high temperature regime (T>450K). These data are consistent with the proposal that monomeric polyglutamine fluctuates between disparate collections of

conformations of roughly equivalent compactness. Intrinsic disorder results because collapse is only weakly coupled from folding in these simple systems that lack the requisite sequence specificity to prefer a specific compact conformation.

IV.4.6. Importance of Spontaneous Fluctuations for Promoting Intermolecular Associations

We have established that monomeric polyglutamine, which is intrinsically disordered, associates to form disordered homodimers. The latter point is underscored in the analysis where we showed that local conformational propensities are essentially unchanged between the isolated monomer and associated dimer, Figure 4.7. Using a simple approach, we interrogated the role of intrinsic disorder (spontaneous fluctuations) of polyglutamine in promoting intermolecular associations. This was done in a series of simulations where we quantified the likelihood of realizing spontaneous associations of rigid globules. These simulations were carried out as follows: random globular conformations were chosen from the conformational ensemble of monomeric Q_{30} at T=298K. The internal coordinates were then frozen and only rigid body Monte Carlo moves were allowed for subsequent sampling. Statistics were recorded to construct the requisite histograms for intermolecular separations sampled in simulations with rigid globules. The process was repeated approximately a thousand times and the resultant, average cumulative distribution $F(R)$ was compared to that obtained for the association of “fully flexible” chains. These comparisons are shown in Figure 4.9:

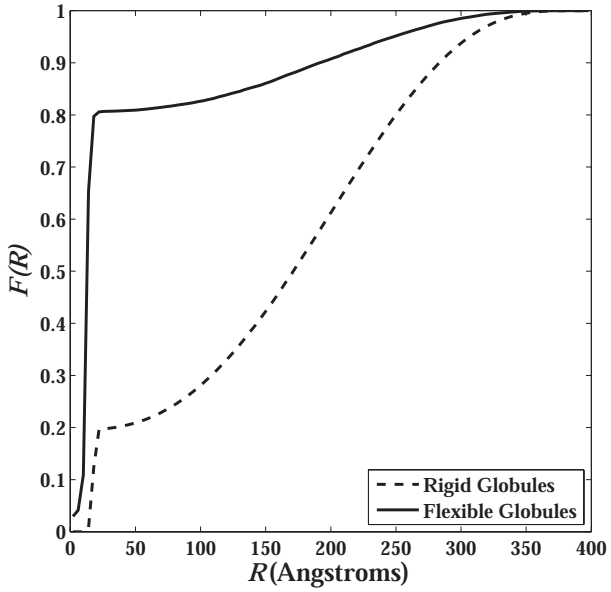

Figure 4.9: **The important of fluctuations for the spontaneity of association**. Cumulative distribution functions $F(R)$ quantifying the probability of realizing an intermolecular separation that is less than or equal to R for pairs of Q_{30} molecules that are either fully flexible (solid curves) or rigid (dashed curves) globules. Data were gathered from Metropolis Monte Carlo simulations that were performed with the simulation temperature T=298K.

The dashed curve, which corresponds to the cumulative distribution function for rigid globules, reveals the importance of conformational fluctuations in promoting intermolecular associations. The suppression of conformational fluctuations at T=298K leads to a diminution of intermolecular associativity for Q_{30}. The lack of rigid structural preferences or a stable fold upon collapse is clearly responsible for promoting intermolecular associations between disordered globules. This result is consistent with the observation that many aggregation sequences are also intrinsically disordered. However, some caution is required in interpreting the results of Figure 4.9. For instance, Figure 4.8 shows that the

degree of disorder measured by σ_N^2 increases with temperature for T<360K and yet $B_{22}(T)$ decreases with increasing temperature. The physical basis for the latter observation comes from the analysis in Figures 4.5 and 4.6, which demonstrates that the poorness of solvent decreases with increasing temperature. Therefore, we conclude that both poorness of solvent and spontaneous conformational fluctuations work together to promote spontaneous homodimerization. This point is reinforced by the following observation: in the temperature regime 360K<T<410K the magnitudes of conformational fluctuations go through a maximum for all chain lengths. In this regime, the surface energy penalty, measured by $C_2(T)$, is still positive and approaches zero only as T approaches T_θ (data not shown). Under these conditions, homodimerization might require the formation of an appropriate conformational nucleus to which only appropriate conformations would be able to dock and minimize the unfavorable interface with the surrounding solvent. Alternatively, some other, higher-order, oligomeric species might be the thermodynamically favored entity because such a species might minimize the unfavorable solute-solvent interface more efficiently than dimers in the regime 360K<T<410K. A detailed investigation of the precise correlation between poorness of solvent and the magnitude of conformational fluctuations merits further scrutiny and is reserved for a separate study.

IV.4.7. Contacts that Promote Collapse and Dimerization

Polyglutamine molecules are polyamides built by a repetition of backbone secondary amides and sidechain primary amides. To analyze the types of inter-

atomic contacts that lead to collapse and dimerization, we computed site-site pair correlation functions. The site-site correlation functions of interest to us are between backbone donors (N) and backbone acceptors (O), sidechain donors (N) and sidechain acceptors (O), backbone donors (N) and sidechain acceptors (O), and sidechain donors (N) and backbone acceptors (O). If we denote donor atoms as D and acceptor atoms as A, then the relevant donor-acceptor site-site correlation function $g_{\mathrm{DA}}(r)$ at temperature T is computed as (see Equation 2-3):

$$g_{\mathrm{DA}}(r) = \frac{h_{\mathrm{DA}}^{(T)}(r)}{h_{\mathrm{DA}}^{(\theta)}(r)} \tag{4-7}$$

Here, $h_{\mathrm{DA}}^{(T)}(r)$ is the histogram of relevant donor-acceptor distances at temperature T and $h_{\mathrm{DA}}^{(\theta)}(r)$ is the corresponding histogram of distances at T_θ. If $g_{\mathrm{DA}}(r) > 1$, then there is an enhancement of the relevant donor-acceptor contacts in the ensemble at temperature T vis-à-vis T_θ; if $g_{\mathrm{DA}}(r) = 1$, then the distribution of donor-acceptor contacts at separation r is equivalent to that of T_θ; finally, if $g_{\mathrm{DA}}(r) < 1$, then there is a depletion of donor-acceptor contacts at separation r vis-à-vis T_θ.

Figure 4.10 shows intramolecular donor-acceptor site-site correlation functions for monomeric Q_{45}. As T approaches T_θ, all pair correlation functions converge upon values of unity for all distances. For lower temperatures, specifically $T \leq 360$K, there is significant enhancement vis-à-vis T_θ of short-range

(3Å≤r<5Å) sidechain donor – sidechain acceptor and sidechain donor – backbone acceptor contacts:

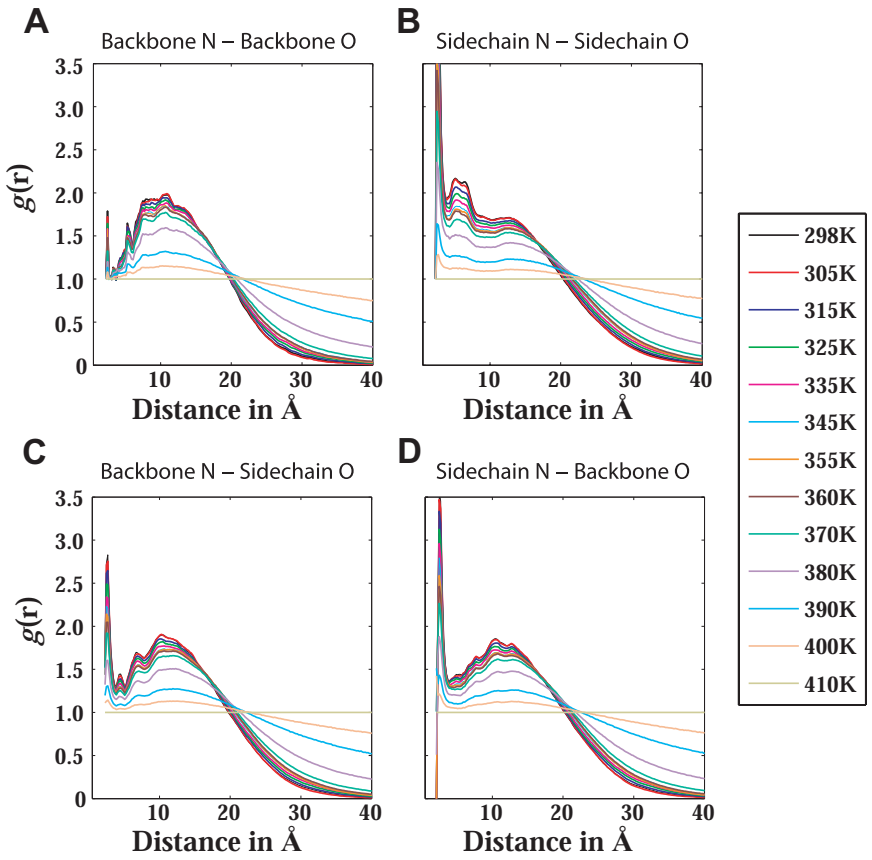

Figure 4.10: **Temperature dependent, *intra*molecular site-site correlation functions for different pairs of backbone and sidechain atoms**. The pair correlation functions were computed using data from simulations of two polyglutamine molecules.

All four sets of site-site correlation functions in Figure 4.10 show systematic enhancements of medium-range contacts (5Å≤r≤20Å) and depletion of distal contacts. This feature is consistent with the preference for collapsed states at lower temperatures. The pair correlation functions shown in Figure 4.10 suggest that the collapsed states are characterized by prominent sidechain-

backbone interactions, again with respect to T_θ, indicating that the sidechain amides solvate the backbone and are thereby minimizing the interface between backbone secondary amides and the aqueous milieu.

The approach used here to calculate pair correlations differs from the one used in Chapter II (see Figure 2.8). Here, we used T_θ as our reference state whereas previously we used an ideal chain model as the reference. Therefore, the two sets of correlation functions for collapsed, monomeric polyglutamine are different. Nonetheless, if we juxtapose the conclusions from Chapter II and here, then in II.4.3 backbone-backbone interactions were identified as the dominant interactions promoting collapse. Conversely, here we identify sidechain-mediated interactions. We can speculate that this result is tied to a fundamental difference between the two models: in ABSINTH, the free energy of solvation for the backbone secondary amide (N-methylacetamide) is set to the experimental value of -10.1 kcal/mol (see Table 3.a). The work in Chapter II, however, employed simulations in explicit solvent using the Tip3p water model[42] and the OPLS-AA/L force field[28]. For this combination of force fields, we measured the free energy of solvation computationally and obtained a value of only -6.5 kcal/mol (Vitalis and Pappu, unpublished), *i.e.*, a value suggesting much less favorable interactions between the backbone model compound and the solvent water. Therefore, it would not seem surprising that backbone-backbone interactions are more important in the work presented in Chapter II. A comparative analysis of mixtures of N-methylacetamide and propionamide (the sidechain model compound) between the two force fields may shed more light on this issue.

Figure 4.11 shows intermolecular donor-acceptor site-site correlation functions calculated using simulation data for a pair of Q_{45} molecules in the simulation volume. These pair correlation functions are shown on a log-scale to facilitate the visualization of all the data:

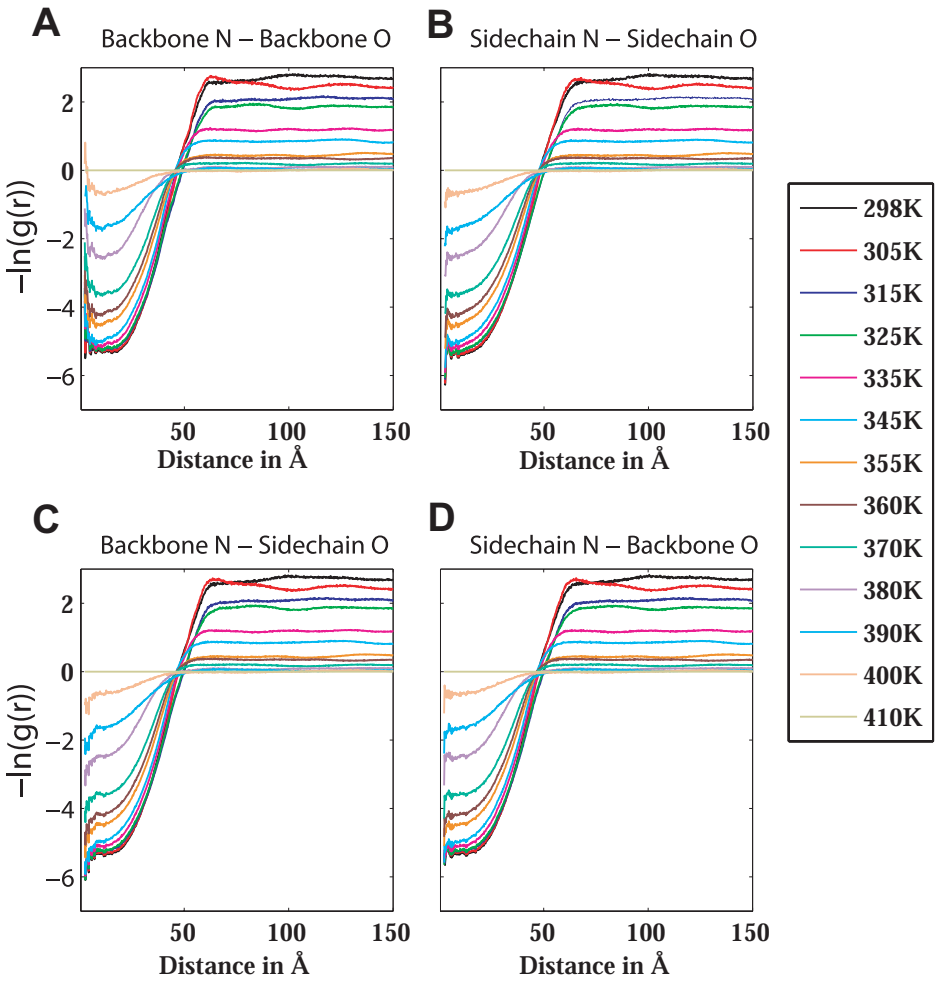

Figure 4.11: **Temperature dependent, *inter*molecular site-site correlation functions for different pairs of backbone and sidechain atoms**. The ordinate is shown in a natural log-scale to facilitate the visualization of the variation of the pair correlation over the entire range of intermolecular separations. As for Figure 4.10, the pair correlation

functions were computed using data from simulations with pairs of polyglutamine molecules.

For temperatures that are in the collapse regime, there is significant enhancement of all flavors of donor-acceptor contacts. This is because of significant intermolecular donor-acceptor contacts that are absent at the theta-point. These observations suggest that spontaneous dimerization of polyglutamine is the result of the drive to minimize the interface with the surrounding aqueous environment, a poor solvent for polyglutamine, and to replace this interface with favorable intra- and intermolecular contacts between all combinations of backbone donors, backbone acceptors, sidechain donors, and sidechain acceptors.

IV.5. Discussion of Implications for the Aggregation of Homopolymeric Polyglutamine

Summary of Results

In this chapter, we have presented results from atomistic simulations on the length and temperature dependence of conformational equilibria and spontaneous dimerization of polyglutamine molecules. Our main findings are as follows:

- Profiles for coil-to-globule transitions (Figure 4.1) show a striking length dependence that is in good agreement with expectations from polymer physics theories. Specifically, the sharpness of this transition increases with N, and this implies that the stability of collapsed states in aqueous milieus

increases with chain length. Curves that plot the variation of normalized chain size with temperature coincide at the “tri-critical” point, T_θ.

- Homodimerization is spontaneous in the collapse regime and its spontaneity increases with chain length (Figure 4.4). Conversely, for a given chain length, its spontaneity decreases with increasing temperature. Homodimerization is tightly coupled to the collapse of monomeric polyglutamine (Figures 4.5 and 4.6).
- Driving forces for collapse of monomeric polyglutamine have two generic components. These are: (i) the drive to maximize self-interactions; and (ii) the drive to minimize the unfavorable solute-solvent interface with the surrounding aqueous milieu. Congruently, dimerization should lead to the formation of additional self-interactions (Figures 4.6 and 4.11) and the number of favorable intra- and intermolecular self-interactions should increase with chain length similar to the “linear lattice” effect proposed by Bennett *et al.*[43] Similarly, dimerization leads to further minimization of the unfavorable solute-solvent interface (Figure 4.6).
- Evidence that the length dependence of spontaneous homodimerization is non-specific and originates in generic considerations for polymers in poor solvents comes from Figure 4.7. Here, we showed that there are no substantial local conformational changes, such as β-sheet formation or conversion from α-helical forms that can be implicated in promoting homodimerization.

- We show, in complete agreement with experimental and computational data,[17,18,23,44] that polyglutamine molecules, irrespective of chain length, are intrinsically disordered even under conditions where collapsed states are thermodynamically favored (Figure 4.8, also compare Chapter II). This intrinsic disorder, *i.e.*, the inability to adopt a stable fold can be implicated in the spontaneity of homodimerization. Suppression of disorder by quenching conformational fluctuations leads to a significant diminution in the preference for associated states (Figure 4.9).
- We have shown that homodimerization is spontaneous and that the spontaneity increases with chain length for a prescribed poorness of the solvent. However, dimerization does not require the obligate formation of a specific, thermodynamically unfavorable, conformational species of the monomeric form (Figures 4.7 and 4.10). Our results also suggest that higher-order oligomers can form readily, although this needs to be studied carefully in future work. Hence, homogeneous nucleation is unlikely to be the correct mechanistic explanation for polyglutamine aggregation.

Connection to Interpretations of Experimental Results

Chen *et al.*[19] described the formation of large ordered polyglutamine aggregates as a nucleation-dependent reaction using the model of Ferrone[36] (see I.2). In the schematic that emerged monomeric polyglutamine (irrespective of chain length) is in rapid pre-equilibrium with an ordered nucleus (presumably an ordered β-sheet conformation). This unfavorable folding reaction is a conformational pre-requisite for the formation of aggregates of all sizes, including

dimers. Four probes were used to monitor the kinetics of aggregation. These were CD, light scattering, ThT binding, and reverse phase HPLC. If β-sheet contents do not vary with oligomerization, then CD signals would not change as oligomers formed. Similarly, light scattering cannot resolve the presence of small oligomers; ThT binding most likely reports only on the formation of large ordered aggregates and is expected to correlate well with the CD signal. Lastly, if the oligomers are part of the soluble species, then reverse phase HPLC would not detect them, either. Hence, the data presented by Chen *et al*.[19] cannot rule out the presence of soluble oligomers in the reacting mixture.

Interestingly, other lines of experimental evidence support the presence of oligomers as identifiable intermediates.[45,46] Recently, Lee *et al.*[47] measured the aggregation kinetics of Q_{23} using peptide constructs that were similar to those used by Chen *et al.*[19] Using both static and dynamic light scattering, these authors found evidence for the formation of soluble, linear aggregates during the lag-phase. They also found the early aggregates to be lacking in regular secondary structure. Inasmuch as we can connect dimer formation with formation of larger aggregates, we propose that our results, which show a lack of local conformational specificity in chain collapse and intermolecular interactions, are consistent with the observations of Lee *et al*. Similar interpretations with respect to the existence of soluble oligomers were obtained *in vivo* by Takahashi *et al*.[48]

The presence of disordered low and high molecular weight aggregates implies that simple homogeneous nucleation models may not accurately describe polyglutamine aggregation. Questions persist regarding the degree of complexity

needed in mechanistic models for polyglutamine aggregation. One possibility is that the formation of disordered linear or spherical aggregates, unlike the formation of ordered aggregates, occurs off-pathway and does not follow the tenets of homogeneous nucleation theory. Alternatively, as suggested by Lee *et al.*[47] and others,[49] disordered aggregates that are sufficiently large might convert to ordered forms. The different scenarios are summarized in Figure 4.12:

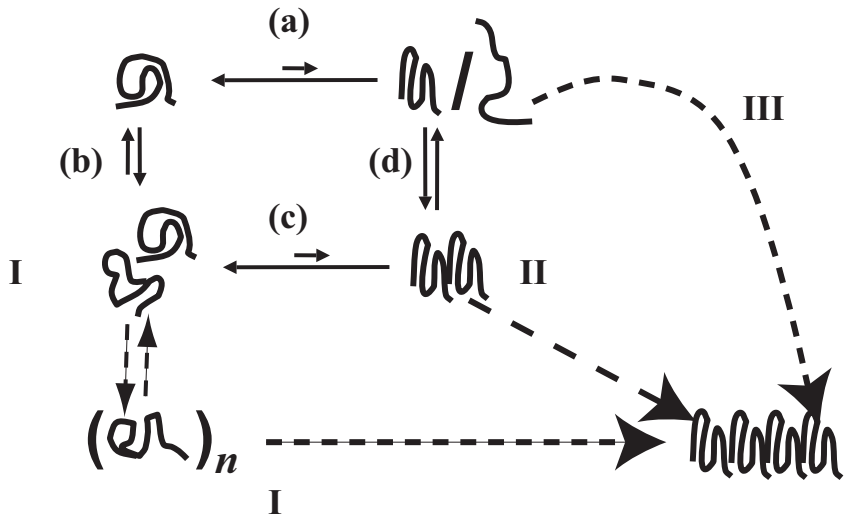

Figure 4.12: Schematic for the formation of higher order aggregates given a prescribed poorness of solvent.

Within the cartoon in Figure 4.12, the simulations in this chapter have addressed steps (a), (b), and (c), albeit at very low copy numbers and high effective concentrations. In step (a) we find that monomer conformational equilibria favor non-specific collapsed states as opposed to a single folded species or an ensemble of extended conformations. Step (b) shows the favorable, spontaneous homodimerization and formation of disordered dimers. Step (c) shows that in our simulations disordered dimers are favored over

ordered dimers. We probe step (d), which requires probing the formation of high energy species along a well-defined reaction coordinate, in Chapter V. The schematic addresses possible routes to the ordered aggregate shown in the bottom right corner of the picture. Scenario I calls for the conformational reorganization within a large, disordered linear / spherical aggregate. This scenario resembles the concept of gelation or ordered aggregation which requires a critical number of intermolecular contacts. Scenario II is that of Chen *et al.*[19] Here, non-specific oligomers represent off-pathway events which effectively modulate the pre-equilibrium implied in (a). Scenario III depicts the formation of ordered aggregates via entanglement of swollen conformations that are also high energy species in a poor solvent.

The schematic is congruent with the tenets of the generalized Lumry-Eyring model put forward recently by Andrews and Roberts.[50] The work in Chapter V will address how likely the different scenarios appear by explicitly probing the impact of structure on dimerization propensity and mechanism. Chapter VI will resolve the impact of sequence context on the qualitative picture depicted in Figure 4.12. This is a crucial point since a fundamental difference between the *in silico* experiments presented here and typical *in vitro* experiments is the ideal hompolymeric nature we stipulate to probe the *intrinsic* properties of glutamine-rich polypeptides. Capped homopolymers are so insoluble, however, that they cannot be studied experimentally and are flanked with charged residues instead.[51]

Alternative Mechanisms Proposed Based on Computer Simulations

Molecular simulations have played an important role in generating insights and testable hypotheses for various self-assembly phenomena involving folded proteins and intrinsically disordered proteins.[52-54] The latter are challenging systems for simulation and experiment alike because their free energy landscapes are both degenerate and rugged. Marchut and Hall[55,56] employed a conceptually different, structurally guided, coarse-grain model to describe peptide and solvent. Despite this difference from our approach, a brief comparison of the results is in order.

In their most recent study,[55] the concentration used is 2.5mM for a system comprised of 24 molecules with chain lengths ranging from 16-48 residues at various reduced temperatures. They find that at temperatures close to the effective T_0 of their model large-scale aggregates with relatively large fractions of β-sheet hydrogen bonds and with distinctive ring-like topologies are populated. The authors note that experimental evidence for these structural motifs is lacking. However, at lower reduced temperatures they describe “amorphous aggregates”, *i.e.*, aggregates lacking in structure. Our results are congruent with this unstructured “phase”. In their parlance, disordered globules with unstructured interfaces are termed amorphous aggregates. As for the observed ordered phase (“sheets”), we argue that the concentration regime in their work as well as the employed model predispose the results toward this order. This is suggested by the fact that β-rich structures appear even at the monomer level, which is incongruent with existing experimental data for monomeric polyglutamine.[17,18] It

is therefore likely that the appearance of an ordered phase for small oligomers is an artifact of the way their models were built. These differences not withstanding, the results of both studies appear to share some overlap in predicting amorphous aggregates under certain conditions. Whether more of the results will be reconciled if the simulation conditions between the two studies are fully matched, remains to be seen, and is a topic for future investigation.

IV.6. Bibliography

1. Vitalis, A.; Wang, X.; Pappu, R. V. J Mol Biol 2008, 384(1), 279-297.
2. Vitalis, A.; Pappu, R. V. J Comput Chem 2009, 30(5), 673-699.
3. Vitalis, A.; Lyle, N.; Pappu, R. V. Biophys J 2009, *in press*.
4. Pappu, R. V.; Wang, X.; Vitalis, A.; Crick, S. L. Arch Biochem Biophys 2007, 469(1), 132-141.
5. Flory, P. J. J Chem Phys 1945, 13(11), 453-465.
6. Flory, P. J. J Chem Phys 1941, 9(8), 660-661.
7. Huggins, M. J. J Phys Chem 1942, 46(1), 151-158.
8. Huggins, M. J. J Chem Phys 1941, 9(5), 440.
9. Chuang, J.; Grosberg, A. Y.; Tanaka, T. J Chem Phys 2000, 112(14), 6434-6442.
10. Raos, G.; Allegra, G. J Chem Phys 1997, 107(16), 6479-6490.
11. Fields, G. B.; Alonso, D. O. V.; Stigter, D.; Dill, K. A. J Phys Chem 1992, 96(10), 3974-3981.
12. Muthukumar, M. J Chem Phys 1986, 85(8), 4722-4728.
13. Grosberg, A. Y.; Kuznetsov, D. V. Macromolecules 1992, 25(7), 1980-1990.
14. Grosberg, A. Y.; Kuznetsov, D. V. Macromolecules 1992, 25(7), 1970-1979.

15. Rabotyagova, O. S.; Cebe, P.; Kaplan, D. L. Biomacromolecules 2009, 10(2), 229-236.

16. Trotter, J. A.; Kadler, K. E.; Holmes, D. F. J Mol Biol 2000, 300(3), 531-540.

17. Masino, L.; Kelly, G.; Leonard, K.; Trottier, Y.; Pastore, A. FEBS Lett 2002, 513(2-3), 267-272.

18. Chen, S.; Berthelier, V.; Yang, W.; Wetzel, R. J Mol Biol 2001, 311(1), 173-182.

19. Chen, S. M.; Ferrone, F. A.; Wetzel, R. Proc Natl Acad Sci USA 2002, 99(18), 11884-11889.

20. Scherzinger, E.; Sittler, A.; Schweiger, K.; Heiser, V.; Lurz, R.; Hasenbank, R.; Bates, G. P.; Lehrach, H.; Wanker, E. E. Proc Natl Acad Sci U S A 1999, 96(8), 4604-4609.

21. Chen, S. M.; Berthelier, V.; Hamilton, J. B.; O'Nuallain, B.; Wetzel, R. Biochemistry 2002, 41(23), 7391-7399.

22. Scherzinger, E.; Lurz, R.; Turmaine, M.; Mangiarini, L.; Hollenbach, B.; Hasenbank, R.; Bates, G. P.; Davies, S. W.; Lehrach, H.; Wanker, E. E. Cell 1997, 90(3), 549-558.

23. Vitalis, A.; Wang, X.; Pappu, R. V. Biophys J 2007, 93(6), 1923-1937.

24. Crick, S. L.; Jayaraman, M.; Frieden, C.; Wetzel, R.; Pappu, R. V. Proc Natl Acad Sci U S A 2006, 103(45), 16764-16769.

25. Bhattacharyya, A. M.; Thakur, A. K.; Wetzel, R. Proc Natl Acad Sci U S A 2005, 102(43), 15400-15405.

26. Vitalis, A.; Steffen, A.; Lyle, N.; Mao, A.; Pappu, R. V. J Chem Theory Comput 2009, *manuscript in preparation*.

27. Sugita, Y.; Okamoto, Y. Chem Phys Lett 1999, 314(1-2), 141-151.

28. Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. J Phys Chem B 2001, 105(28), 6474-6487.

29. Imbert, J. B.; Lesne, A.; Victor, J. M. Phys Rev E 1997, 56(5), 5630-5647.

30. Yang, S.; Levine, H.; Onuchic, J. N. J Mol Biol 2005, 352(1), 202-211.

31. Withers, I. M.; Dobrynin, A. V.; Berkowitz, M. L.; Rubinstein, M. J Chem Phys 2003, 118(10), 4721-4732.

32. Milchev, A.; Paul, W.; Binder, K. J Chem Phys 1993, 99(6), 4786-4798.

33. Dobson, C. M.; Swoboda, B. E. P.; Joniau, M.; Weissman, C. Philos Trans R Soc Lond B Biol Sci 2001, 356(1406), 133-145.

34. Walker, F. O. Lancet 2007, 369(9557), 218-228.

35. Talanquer, V.; Oxtoby, D. W. J Chem Phys 1998, 109(1), 223-227.

36. Ferrone, F. In Amyloid, Prions, And Other Protein Aggregates, 1999, p 256-274.

37. Kabsch, W.; Sander, C. Biopolymers - Peptide Science Section 1983, 22(12), 2577-2637.

38. Krishnan, R.; Lindquist, S. L. Nature 2005, 435(7043), 765-772.

39. Derkatch, I. L.; Uptain, S. M.; Outeiro, T. F.; Krishnan, R.; Lindquist, S. L.; Liebman, S. W. Proc Natl Acad Sci U S A 2004, 101(35), 12934-12939.

40. Mukhopadhyay, S.; Krishnan, R.; Lemke, E. A.; Lindquist, S.; Deniz, A. A. Proc Natl Acad Sci U S A 2007, 104(8), 2649-2654.

41. Enderlein, J. ChemPhysChem 2007, 8(11), 1607-1609.

42. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. J Chem Phys 1983, 79(2), 926-935.

43. Bennett, M. J.; Huey-Tubman, K. E.; Herr, A. B.; West, A. P.; Ross, S. A.; Bjorkman, P. J. Proc Natl Acad Sci U S A 2002, 99(18), 11634-11639.

44. Wang, X. L.; Vitalis, A.; Wyczalkowski, M. A.; Pappu, R. V. Prot Struct Funct Bioinf 2006, 63(2), 297-311.

45. Mukai, H.; Isagawa, T.; Goyama, E.; Tanaka, S.; Bence, N. F.; Tamura, A.; Ono, Y.; Kopito, R. R. Proc Natl Acad Sci U S A 2005, 102(31), 10887-10892.

46. Wacker, J. L.; Zareie, M. H.; Fong, H.; Sarikaya, M.; Muchowski, P. J. Nat Struct Mol Biol 2004, 11(12), 1215-1222.

47. Lee, C. C.; Walters, R. H.; Murphy, R. M. Biochemistry 2007, 46(44), 12810-12820.

48. Takahashi, T.; Kikuchi, S.; Katada, S.; Nagai, Y.; Nishizawa, M.; Onodera, O. Hum Mol Genet 2008, 17(3), 345-356.

49. Poirier, M. A.; Li, H. L.; Macosko, J.; Cai, S. W.; Amzel, M.; Ross, C. A. J Biol Chem 2002, 277(43), 41032-41037.

50. Andrews, J. M.; Roberts, C. J. J Phys Chem B 2007, 111(27), 7897-7913.

51. Berthelier, V.; Wetzel, R. Methods in molecular biology (Clifton, NJ) 2003, 217, 295-303.

52. Bellesia, G.; Shea, J. E. J Chem Phys 2007, 126(24), art. no. 245104.

53. Pellarin, R.; Caflisch, A. J Mol Biol 2006, 360(4), 882-892.

54. Tarus, B.; Straub, J. E.; Thirumalai, D. J Mol Biol 2005, 345(5), 1141-1156.

55. Marchut, A. J.; Hall, C. K. Prot Struct Funct Bioinf 2007, 66(1), 96-109.

56. Marchut, A. J.; Hall, C. K. Biophys J 2006, 90(12), 4574-4584.

CHAPTER V. THE THERMODYNAMICS OF β-SHEET FORMATION FOR MONOMERIC AND DIMERIC POLYGLUTAMINE

V.1. Preamble

One of the prominent results of Chapter IV[1] is that structure does not seem to be a requirement for two homopolymers composed of glutamine to favorably interact at concentrations in the μM-range. Such a generic propensity to dimerize would suggest that larger oligomers form readily and are spherical and disordered in nature. One proposal would be that canonical β-secondary structure can much more readily form in a water-deprived ("dry") environment and that larger assemblies are hence needed to observe conformational re-arrangement toward higher β-content. Such a simulation would be a computationally infeasible endeavor with current resources and remains reserved for future work. Even then, significant effort might have to be spent upfront to coarse-grain the representation of the system further (see Chapter VII).

In silico, we do enjoy the benefit of having access to tools which allow the (biased) sampling of low likelihood species. When set up properly, such an "unphysical" simulation can reliably yield thermodynamic data for regions of phase space which would not be visited during a finite length equilibrium simulation. We take full advantage of this idea in this chapter. Nicholas Lyle, the co-author on the manuscript underlying this chapter,[2] took over an established protocol for biased simulations quantifying the thermodynamics of the formation

of β-secondary structure in polyglutamine. He also adopted the technology and analyses introduced in Chapter IV to quantify the associativity of two polyglutamine chains. His contributions to the work in this chapter are as follows: he verified that the reaction coordinate which measures β-content is meaningful and fine-tuned its parameters. All the data presented in this chapter were generated by him. He analyzed those data and created the figures based on suggestions and specific requests.

V.2. Introduction to the Putative Role of β-Secondary Structure during the Early Stages of Polyglutamine Aggregation

As was outlined in I.2.6, one of the key findings in the molecular characterization of CAG repeat diseases was the discovery that aggregates rich in polyglutamine display amyloid-like features, *i.e.*, that they trigger the characteristic fluorescence shift upon ThT binding and exhibit fibrillar architectures in electron micrographs.[3,4] CD spectroscopy adjudicates polyglutamine-based aggregates to be rich in β-secondary structure, which we have to consider agnostically: β-helices,[5] parallel and antiparallel β-sheets,[6] and β-hairpins are all putative secondary structure motifs giving rise to the characteristic CD signal.

Such considerations appear to place polyQ-aggregates firmly in the realm of amyloids. Short sequences derived from amyloidogenic proteins including the peptide GNNQQNY from the yeast prion Sup35 have been successfully crystallized and studied by X-ray diffraction.[7,8] The structure identified reveals a

cross β-spine architecture and suggests β-secondary structure as the dominant conformation in condensed peptide phases. It is difficult, however, to obtain such microcrystals consistently,[9] and their relevance as reporters even for the structure of fibrillar aggregates of amyloidogenic proteins may be questioned.

Solid-state NMR data have been used to derive structural models for fibrils formed by the most prominent amyloidogenic peptide, *i.e.*, the Alzheimer's peptide Aβ. These models provide the only atomistic structures of directly disease-related amyloids. Interestingly, polymorphism is obtained even when fibrils are grown under identical conditions.[10] Much like polyQ-based peptides, Aβ is intrinsically disordered but known to exhibit some transient structural preferences and to form oligomers of specific size.[11,12] Neither of those is true for polyglutamine. We might expect based on the complete lack of specificity (see Chapters II[13] and IV) that polymorphism in aggregates of polyQ-based peptides is amplified. However, to our knowledge, no comprehensive analysis of this has been performed to date. We speculate that it is not at all clear that a heterogeneous environment will consistently yield amyloid-like aggregates for polyglutamine, in particular *in vivo*. In fact, a hypothesis may be formulated that under a wide range of conditions amorphous aggregates are the thermodynamically stable phase or at least represent a necessary reaction intermediate for fibril formation.[14-16] Of course, amorphous aggregates are ill-defined and hence difficult to quantify using the aforementioned, experimental techniques.

The above discussion does not suggest any particular relevance of β-secondary structure during the *early* stages of polyglutamine aggregation. As detailed in I.2.6, monomeric, soluble peptides rich in polyglutamine are usually completely disordered, even though a toxic monomeric conformation rich in β-sheet has been identified for a fusion construct with a non-native host protein.[17] As a first-pass model, Wetzel and co-workers[18,19] have suggested an aggregation mechanism involving a toxic folding event[20] at the monomer level to yield a β-rich nucleus (see I.2.5 and Figure 1.3). It has been argued that the intrinsic disorder and hence its ability to undergo large-scale conformational transitions predispose polyglutamine to such a mechanism.[21] Evidence against a mechanism like this comes from studies in which soluble intermediates were discovered and characterized as being free of canonical secondary structure.[22,23] This would be much more consistent with the proposals brought forth in IV.5 and Figure 4.12.

Can these observations be reconciled? Three major tenets of the model in Figure 1.3 are the assumption of homogeneous nucleation, the nucleus size of one, and the proposal that the conformation of the nucleus is the same as in the final aggregate. It is possible that the observed disordered and soluble oligomers occur off-pathway and that their presence gives rise to an effective nucleation rate by depleting the monomer pool. This would allow aggregation to remain homogeneously nucleated by a monomeric conformer rich in β-secondary structure. On the other hand, it seems quite plausible that the physical

mechanism of aggregation is much more heterogeneous and that the details are masked by the simplicity of the kinetic analysis (see I.2.6).[24-28]

In summary, the mechanistic importance of β-secondary structure during the early stages of polyglutamine aggregation remains unknown. This is primarily due to the difficulty in experimentally characterizing the involved species.[29] The prevalence of this conformation of the peptide backbone in condensed phases deprived of water appears to support the idea that β-rich conformers will self-interact more favorably than conformers with low β-content. We may also conjecture that monomeric, β-rich structures are difficult to populate for short homopolymers since α-helical and disordered conformations might be better suited to sequester the protein backbone from the solvent and to satisfy the collapse constraint imposed upon all uncharged and polar (and not just hydrophobic) polypeptides.[13,30] In this chapter, we test both of those conjectures directly. Specifically we ask:

- How likely is the formation of an ordered, β-rich structure at the monomer level? Are such species observed as stable or metastable states along a reaction coordinate measuring the net β-content of the chain?
- Does the likelihood of forming species high in β-content increase, decrease, or stay constant with increasing chain length? Does the result support the model that the chain length dependence of the rate aggregation is explained by more favorable nucleation with increasing chain length (see I.2.5)?

- Do ordered species increase, reduce, or do they not affect the spontaneous driving force for dimer formation of polyglutamine peptides?
- Are the structural signatures associated with the dimerization of polyglutamine altered if high β-content is enforced, in particular at the dimerization interface?

We organize the remainders of this chapter as follows: first, the necessary details of our methodology are introduced. Emphasis is given to parts that differ from the work presented in Chapter IV (see IV.3). We then show data to justify our approach and provide answers to all of the above questions. We conclude with a summary and discussion of our results in the context of the structural aspects of the early stages of polyglutamine aggregation.

V.3. Simulation Details

V.3.1. The Reaction Coordinate f_β

In Chapter IV, we have defined global metrics of secondary structure content (see Equation 4-5). Here we will employ this definition to define a reaction coordinate f_β that allows quantification of global β-content. As a reminder, f_β relies on measuring the fraction of residues whose ϕ,ψ-angles occupy the region in ϕ,ψ-space characterized as the β-basin:

$$f_\beta = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}1.0 & \text{if } (\phi_i,\psi_i)\in\beta \\ \exp\left(-\tau_\beta d_{(i)}^2\right) & \text{otherwise}\end{cases}$$
$$d_{(i)} = \left\{\left(\sqrt{\left[\left(\phi_i-\phi_\beta\right)\bmod 2\pi\right]^2+\left[\left(\psi_i-\psi_\beta\right)\bmod 2\pi\right]^2-r_\beta}\right)\bmod 2\pi\right\} \quad (5\text{-}1)$$

In Equation 5-1, f_β measures the fractional β-content of the polypeptide chain by averaging over the N residues with polypeptide ϕ,ψ-angles. The $\mathrm{mod}2\pi$ terms correct for periodicity effects when calculating distances in angular space. The coordinates (ϕ_β,ψ_β) define the reference ϕ,ψ-values to be adopted by an individual residue. If they lie within a circle of radius r_β, then they contribute a full fractional count $1/N$ to the total β-content; otherwise, residue i will contribute a reduced fractional count in the interval $[0{:}1/N]$. Its precise value is determined by two parameters, namely the angular distance $d_{(i)}$ and the decay parameter τ_β. The latter is the width of the Gaussian function used to determine the value to be assigned for f_β and ensures a continuously differentiable function.

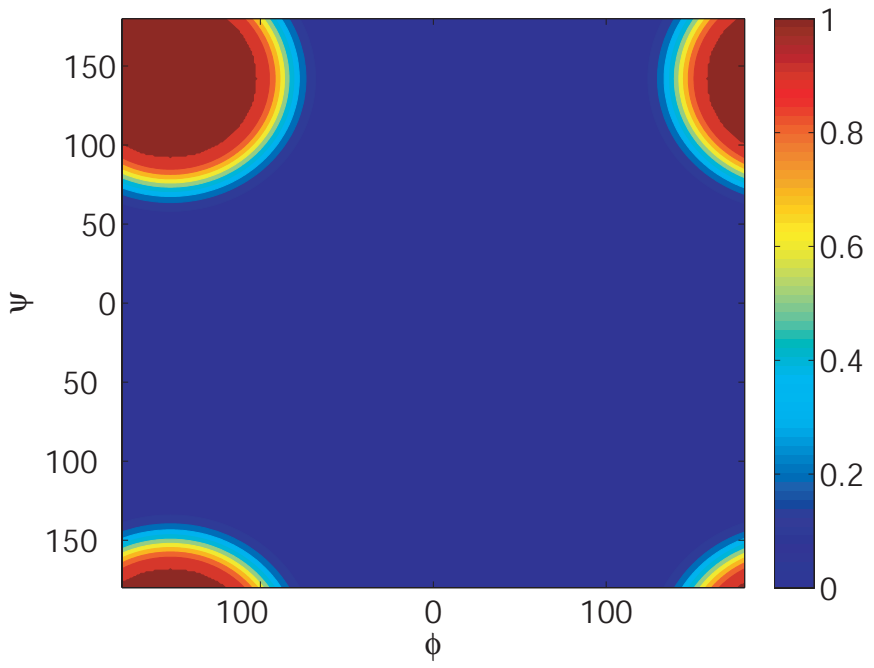

Figure 5.1: **The reaction coordinate f_β as a function of the ϕ,ψ-angles**. Parameter values for r_β and τ_β are as defined in text.

The parameters were obtained via a calibration process: $(\phi_\beta,\psi_\beta)=(-152°,142°)$, $r_\beta=50°$, and $\tau_\beta=0.002\text{deg}^{-2}$. The resultant profile for a single residue in ϕ,ψ-space is shown in Figure 5.1. Data supporting our choices are reported in V.4.1.

V.3.2. Biased conformational Sampling and System Setup

The basic approach to conformational sampling via MMC simulations of the dihedral angles of the peptides was identical to the work in Chapter IV (see Table 4.a). For clarity, Table 5.a summarizes details of the move set employed in this chapter:

Move type	**Settings for simulations of monomeric polyglutamine**	**Settings for simulations with pairs of polyglutamine molecules**
Rigid-body	0%	30% (50%, 10Å, 20°)
Omega (ω)	7% (90%, 5°)	4.9% (90%, 5°)
Sidechain (χ_1, χ_2, χ_3)	30% (4x, 60%, 30°)	21% (4x, 60%, 30°)
Backbone ϕ/ψ	63% (70%, 10°)	44.1% (70%, 10°)

Table 5.a: Overview of the frequency of the different Monte Carlo moves sets used in biased simulations of monomeric and pairs of polyglutamine molecules. Please refer to the caption to Table 4.a for details.

The starting conformations for all simulations of either single polyglutamine chains or pairs of polyglutamine chains were 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. A

droplet boundary condition was used in all cases (Equation 4-1). For monomer simulations, the concentration corresponded to infinite dilution conditions, and for dimer simulations it was 100μM. The ABSINTH model was used as presented in Chapter III (including the parameter settings given in Tables 3.a, 3.b and 3.d). The small modification to the LJ parameters detailed in IV.3 is *not* present here. This means that the results show small differences for the cases in which a direct comparison is possible (see V.4.5 and V.4.6 in particular).

For the majority of the work in this chapter, Equation 3-1 is augmented by and additional potential energy term which restrains the reaction coordinate f_β via a harmonic umbrella potential:[31]

$$U_{f_\beta} = k_\beta \cdot \left(f_\beta - f_\beta^0 \right)^2 \qquad (5\text{-}2)$$

Here, k_β is the spring constant determining the stiffness of the potential, and f_β^0 is the equilibrium position of the restraint. The restraint potential has two advantages: i) it allows us to map out the phase diagram of polyglutamine with the two axes being general solvent quality modulated by simulation temperature (see Chapter IV) and β-content modulated by k_β and f_β^0; and ii) it allows us to employ ensemble re-weighting techniques (WHAM)[32] to obtain free energy profiles along f_β even if the adopted values correspond to extremely low likelihood regions of phase space.

In order to determine the free energy profile along f_β under poor solvent conditions, we performed simulations of monomeric polyglutamine at 298K. For each chain length, we performed eleven sets of distinct umbrella sampling

simulations and in each simulation f_β was restrained to one of eleven target f_β^0-values: [0.0, 0.1, 0.25, 0.3, 0.4, 0.5, 0.6, 0.75, 0.8, 0.9, 1.0]. This initial coarse schedule was augmented by additional simulations to test the robustness of the analysis: based on overlap statistics, six additional values for f_β^0 were considered: [0.2, 0.35, 0.45, 0.55, 0.7, 0.95]. Due to the lower computational cost for these systems, we repeated the REX umbrella sampling calculation for Q_5 and Q_{15} with the full schedule comprised of 17 independent values for f_β^0 and found the results to be independent of schedule density (data not shown). All analyses reported in V.4.3 employ the full set of 17 replicas in the WHAM reconstruction. The number of ϕ,ψ-pairs that contribute to f_β increases with N and the value of k_β in Equation 5-2 varied with N. We used values of k_β=25kcal/mol and 75kcal/mol for Q_5 and Q_{15}, and k_β=150kcal/mol for Q_{30} and Q_{45}, respectively. Therefore, k_β varies from 1.7kcal/mol (Q_{45}) to 2.5kcal/mol (Q_5, Q_{15}, and Q_{30}) per restrained degree of freedom. For each window, sampling was enhanced using the REX technique in f_β-space.[33] In contrast to Chapter IV, swaps were only allowed between neighboring replicas. For each chain length, we performed three independent REX umbrella sampling MC runs using the coarse schedule provided above. 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 (see V.4.1).

For polyglutamine dimer simulations, we combined MC simulations with thermal REX. 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 (see V.4.5). We wish to quantify the spontaneity of intermolecular associations in the poor solvent regime. However, 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 $T_\theta \approx 390$K to ensure that the replicas were used judiciously.

Lastly, error analysis proceeded in identical fashion to Chapter IV (see IV.3.4).

V.4. Results

This section is structured such that the biologically relevant results are presented coherently from V.4.3 onward. V.4.1 and V.4.2 are sections demonstrating the validity of the approach taken in this chapter, which might not appear as intuitive as – for example – the work presented in Chapter IV.

V.4.1. Robustness of Data from Restrained Simulations

Since the use of a restraint potential on f_β is unprecedented, here we strive to provide evidence that the results are reproducible. Figure 5.2 shows the

histograms obtained for the reaction coordinate in the biased simulations of monomeric polyglutamine at 298K:

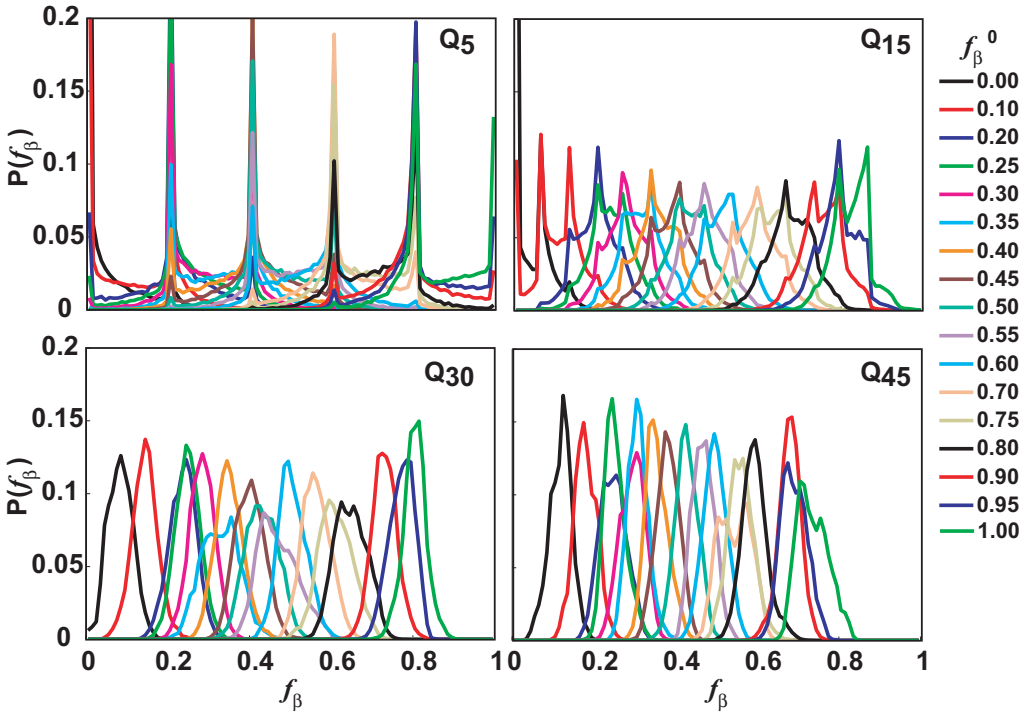

Figure 5.2: Biased histograms of f_β. Data were obtained using a Hamiltonian augmented by Equation 5-2 and for values of f_β^0 as indicated in the figure legend. The plots for Q_5 and Q_{15} reveal the underlying discrete nature of the reaction coordinate. In the limit of τ_β approaching infinity, f_β becomes a strictly discrete function which only adopts fractional values corresponding to the number of residues in the β-basin: e.g., for Q_5 possible values would be 0/5, 1/5, 2/5, 3/5, 4/5, and 5/5 – the longer the chain, the smoother the histograms.

Figure 5.2 reveals that – while there does seem to be some amount of noise in the data for longer chain lengths – the overlap between neighboring replicas X and $X+1$ is generally quite high with the complete schedule of f_β^0-values. It can be quantified using an overlap measure $O_{X,X+1}$:

$$O_{X,X+1} = \frac{2 - \int_{f_\beta=0}^{f_\beta=1} \left| P_X\left(f_\beta\right) - P_{X+1}\left(f_\beta\right) \right| df_\beta}{2} \qquad (5\text{-}3)$$

Here, $P_X(f_\beta)$ is the probability of observing a specific f_β-value in replica X which is characterized by a specific f_β^0-value. Figure 5.3 shows a plot of $O_{X,X+1}$ for all restrained sets of simulations of monomeric polyglutamine:

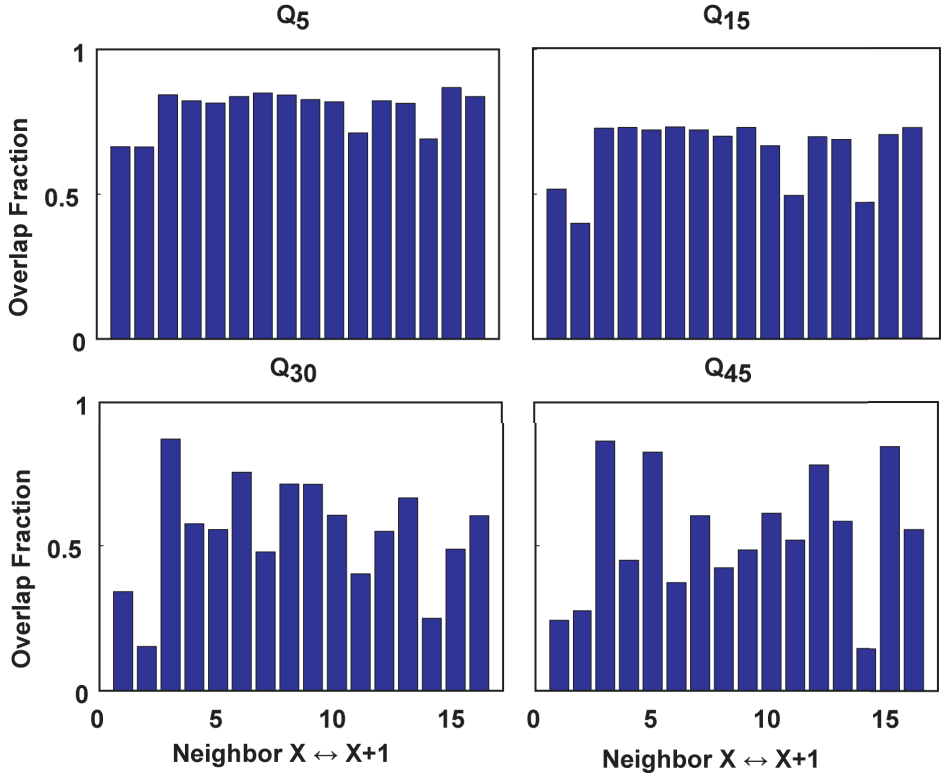

Figure 5.3: Overlap statistics for simulations with restraints on f_β. Overlap metrics between neighboring replicas as calculated by Equation 5-3 are shown as bar plots for simulations of monomeric polyglutamine for all chain lengths.

Figure 5.3 suggests that the overlap between neighboring replicas is sufficient to ensure reliable data and an efficient REX-protocol (see V.3.2). As a final test of the robustness of the analysis, we show that two independent methods of

quantifying the free energy differences between adjacent f_β^0-values give the same results, and – furthermore – that those results are independent of whether a coarser or a finer schedule for f_β^0 is used:

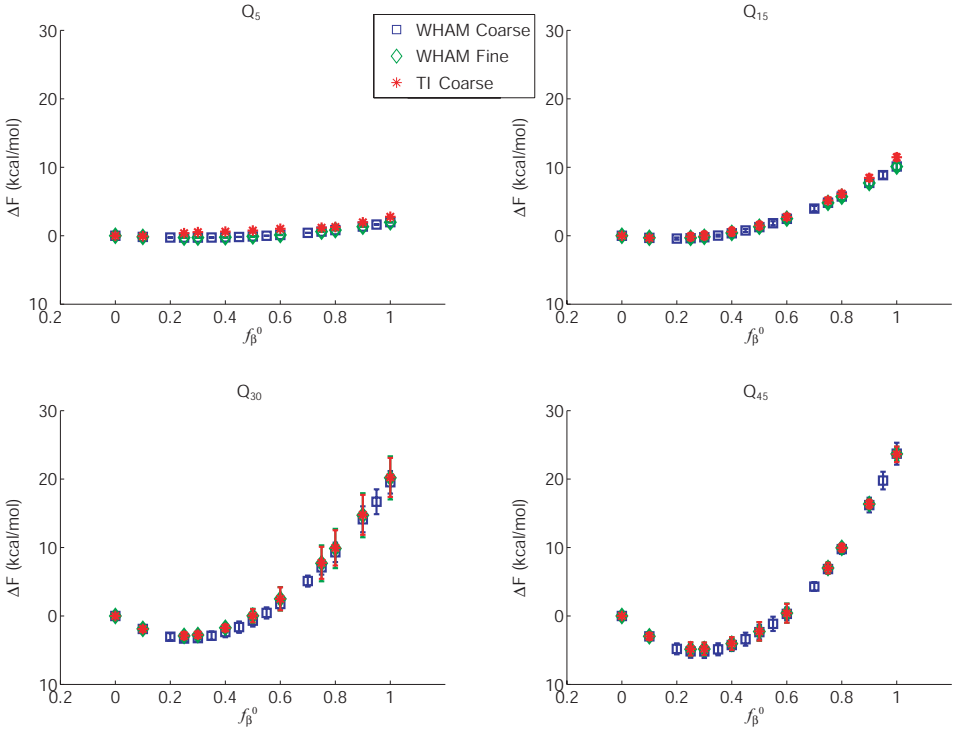

Figure 5.4: Free energy differences between adjacent f_β^0-windows obtained using independent methods. A comparison between the cumulative free energies between adjacent windows is shown. Data were obtained using the iterative WHAM procedure or thermodynamic integration (TI), and are tested for robustness when using a sparser schedule of f_β^0-values.

Figure 5.4 demonstrates that the free energy differences between neighboring f_β^0-values are independent of both the methodology employed to obtain them (TI or WHAM) and of whether the six additional f_β^0-values, which are

part of the finer schedule, are included or not. We therefore conclude that the analysis below is robust and reproducible.

V.4.2. Validity of f_β as a Reaction Coordinate

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[34] 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,[35] normalized by the number of residues, as an alternative to measure the degree of ordered β-sheet. DSSP E-scores are entirely based on hydrogen bond patterns and therefore represent an excellent complementary measure to f_β which is based entirely on dihedral angle populations.

Figure 5.5 shows the correlation between f_β and fractional DSSP E-scores. The correlation coefficient is 0.83 between the two independent metrics. As detailed in the caption to Figure 5.5, E-scores may be zero for structures free of any canonical β-hydrogen bonds whereas f_β never rigorously approaches zero. The differing stringencies for the two criteria give rise to scatter in Figure 5.5. Figure 5.6 shows ribbon drawings of three-dimensional structures for five structures from the database to illustrate that such scatter is easily explainable by the type of structure present.

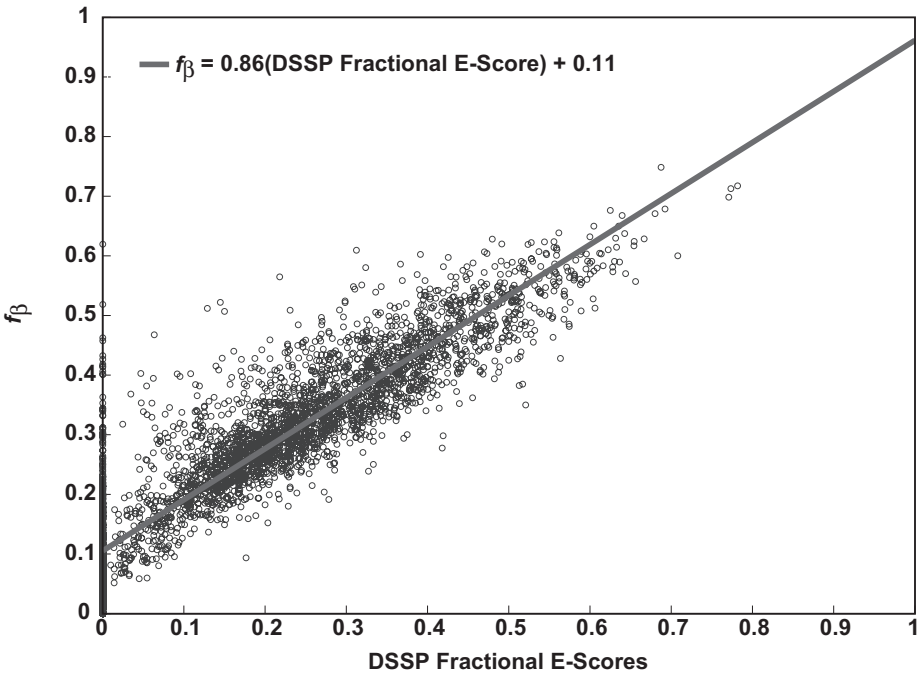

Figure 5.5: Correlation between f_β and fractional DSSP E-scores. The solid line is the line of best fit that quantifies the strength and direction of the linear correlation between f_β values and fractional DSSP E-scores. Parameters for the slope and intercept are shown in the inset. Structures that 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.

Since we do not have definitive prior knowledge of the type of ordered β-sheets that polyglutamine molecules adopt in fibrillar aggregates, it appears reasonable to use a reaction coordinate which reflects that degeneracy, *i.e.*, f_β, instead of the more stringent fractional DSSP E-scores. The high degree of correlation shows that nonetheless f_β is an informative readout of net β-content.

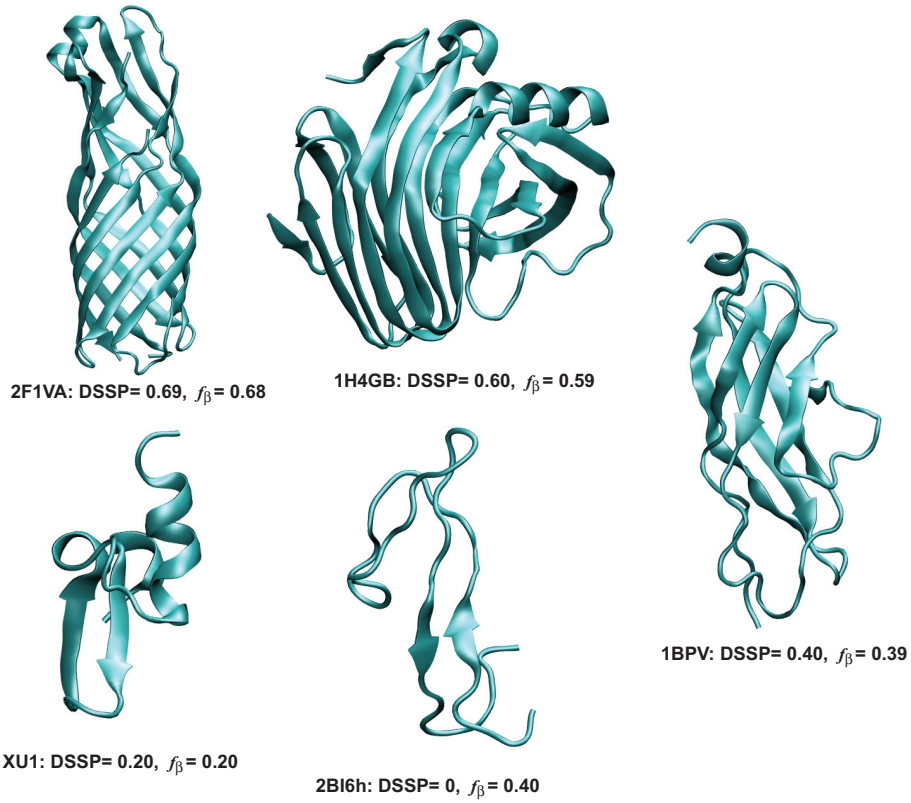

Figure 5.6: Ribbon diagram illustrations of the correlation between f_β and fractional DSSP E-scores. A prominent outlier is shown in 2BI6H which adopts mostly extended structures giving rise to a non-zero f_β-value but has no canonical β-hydrogen bonds as defined by DSSP. Graphics were generated using VMD.[36]

V.4.3. Potentials of Mean Force as a Function of f_β for Monomeric Polyglutamine

The first question of interest is the likelihood of forming β-rich structures at the monomer level for polyglutamine. As outlined in V.2, this emerges directly from proposals in which the formation of a rare, β-rich species represents the nucleation event for polyglutamine aggregation. From simulations employing restraints on a suitable reaction coordinate (see V.3.2, V.4.1 and V.4.2), the free

energy profile along an axis measuring β-content, *viz*. f_β, can be obtained using WHAM. Figure 5.7 shows our results:

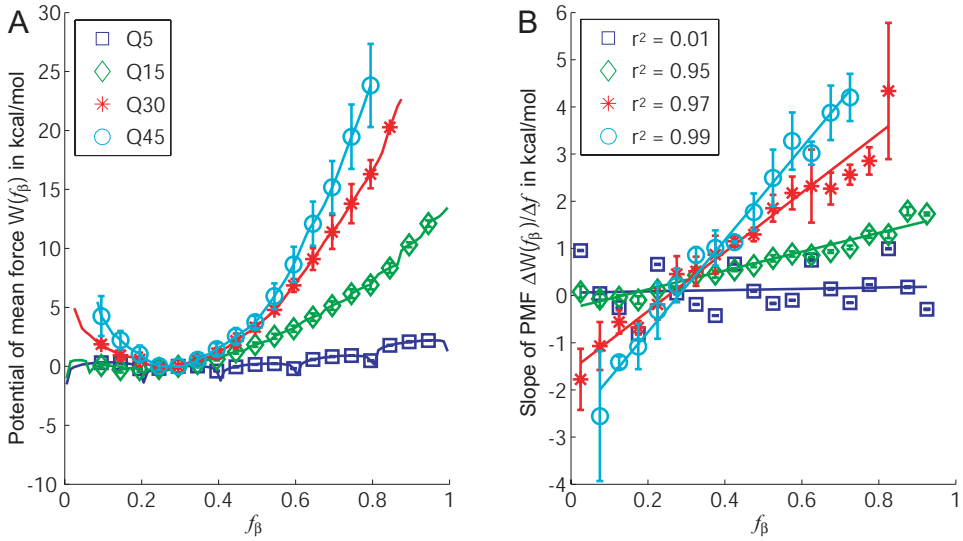

Figure 5.7: Free energy profiles along the reaction coordinate f_β. Panel A shows the PMFs with standard errors. Panel B shows the finite differences of the PMFs, *i.e.*, the mean force. The insets show the Pearson correlation coefficients – r^2 – that diagnose the strength and direction of the hypothesis that the mean force profiles 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). Both slopes and intercepts have units of kcal/mol.

Panel A of Figure 5.7 shows free energy profiles for monomeric polylgutamine of four chain lengths at 298K. The profiles along f_β are mostly featureless, and possess broad minima for $N \geq 15$. For Q_5 the PMF is almost entirely flat and this is consistent with the idea that this short peptide preferentially populates extended conformations. The minima for longer chains are all located at small to intermediate values of f_β, consistently around 0.3. The free energy penalties for accessing conformations with values of f_β that deviate

significantly from 0.3 are large and this is especially true for high values of f_β. For $N \geq 15$, polyglutamine generally prefers compact, disordered states because water is a poor solvent for these polymers.[37] At low temperatures, this is true irrespective of the presence / absence of restraints on f_β. There is no inherent driving force for secondary structure formation with the exception of a somewhat pronounced α-helical propensity for Q_{15} (data not shown). 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, since the α-helix represents a well-populated, favorable conformation for which f_β equals zero. The decreased α-helical propensity for longer chain lengths leads to an increase in the free energy penalty for $f_\beta < 0.3$. The results in Chapter IV exhibit the same trend in helix propensity with chain length (see IV.4.4).[1]

The free energy profiles do not show evidence for distinct, local minima, which would be indicative of the presence of metastable states. This is confirmed by analyzing the derivatives of the PMFs. In the harmonic limit with a single minimum a mean force profile will be a straight line. This is what we observe as in Panel B of Figure 5.7. Absence of anharmonicities supports the conclusion that monomeric polyglutamine does not have access to metastable, β-rich states in isolation. While this result questions the term “pre-equilibrium” used in the kinetic analysis (see I.2.5, in particular Equation 1-2), it does not render the computational and experimentally data mutually inconsistent.

Bhattacharya *et al.*[19] have estimated the nucleation free energy for Q_{47} to be roughly 12.2kcal/mol. Following the analysis shown in Figures 5.5 and 5.6, we expect monomeric, β-rich putative nuclei to correspond to values of f_β between 0.6 and 0.7, *i.e.*, significantly less than unity. This is due to the topological requirement for turn and loop formation in canonical all-β folds. From Panel A in Figure 5.7 we estimate that such structures are associated with free energy penalties of 10-15kcal/mol from the ground state for Q_{45}. This estimated range is consistent with the result obtained by Bhattacharya *et al.*

Next, we turn our attention to the chain length dependence of the PMFs. Wetzel and coworkers proposed a model in which the aggregation rate dependence on chain length is tied to an increased nucleation rate, *i.e.*, a reduced free energy barrier for forming the monomeric, β-rich nucleus.[19] Our data as shown in Figure 5.7 suggest an opposite trend since for four chain lengths spanning the *in vivo* threshold range (N≈37 for HD and similar for several other CAG repeat diseases)[38] the penalty associated with forming structures with high f_β increases monotonically. As explained below (see V.4.7), we do provide evidence which supports the belief that high levels of β-content are a property of larger aggregates. Our calculated free energy profiles do not support the hypothesis that the spontaneity of β-secondary structure formation increases with increasing chain length and thereby contradict the proposal of Chen *et al*.[18] The data *do* reveal that the increase in barrier height appears to level off for Q_{45}, *i.e.*, the PMFs for Q_{30} and Q_{45} bear much more resemblance to each other than those

for Q_{30} and Q_{15}. This observation suggests the possibility of a disappearance or even reversal of the trend of increasing free energy penalty with increasing chain length. Such a result is unlikely going to reconcile our observations with the hypotheses of Chen *et al*., however, since experimentally the aggregation rate increases *monotonically* with chain length.[18]

V.4.4. Structural Characterization of Monomers with High β-Content

In Figure 5.5 (see V.4.2) we show that β-content can be measured by two independent parameters, *viz.* DSSP E-scores and f_β, and that the assignments correlate well. In Figure 5.8 we present a scatter plot of the observed f_β-values in the saved snapshots of *all* simulations for monomeric polyglutamine of chain length $N \geq 15$ against the DSSP E-scores calculated for those conformations. Here, the data are simply pooled and the results illustrate the accessible range of values for one measure β-content given a prescribed value for the other.

Figure 5.8 makes two important points. First, there is good correlation between f_β and the maximally accessible DSSP E-scores which are simply the fractions of all residues assigned to be part of canonical β-secondary structure. This implies that no structures are observed in which the value of f_β is misleading, *i.e.*, in which it would predict low β-content in the presence of characteristic β-hydrogen bond patterns. Second, there is a very substantial spread in the observed E-scores which indicates heterogeneity, *i.e.*, the presence of disordered conformations with no consistent backbone-backbone hydrogen bond patterns even though the value for f_β may be high.

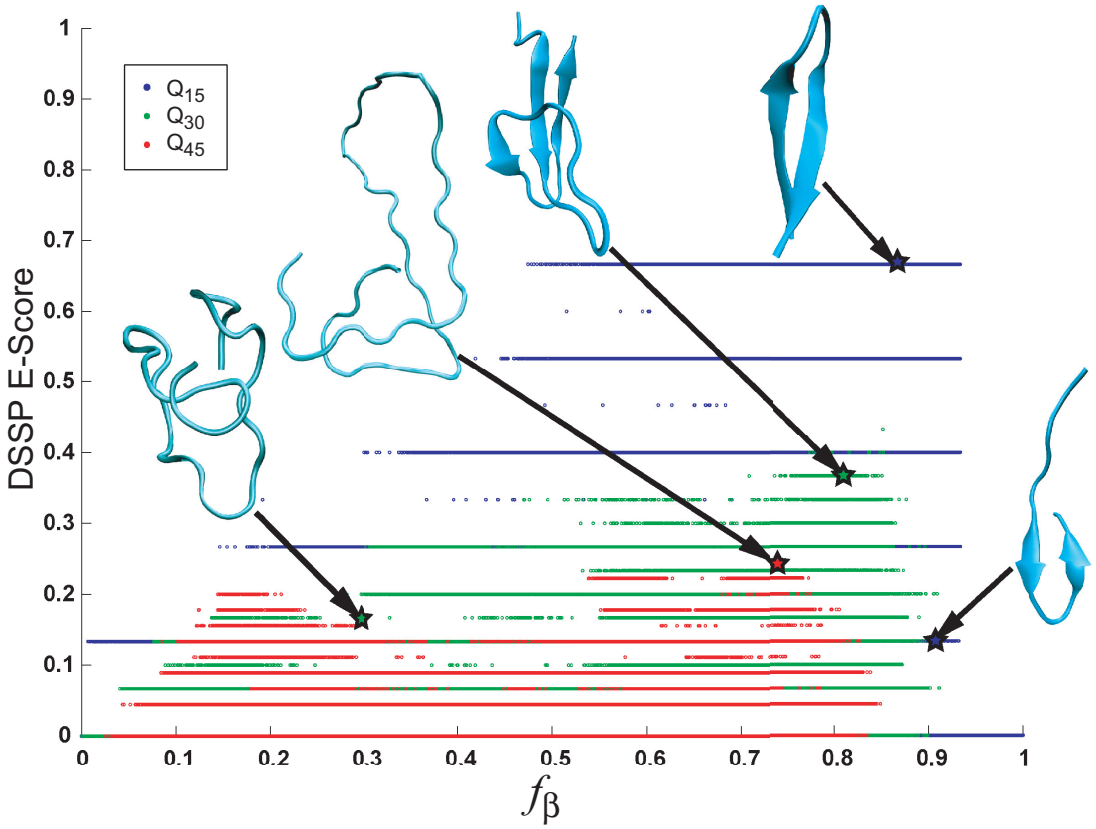

Figure 5.8: Scatter plot of DSSP E-scores and f_β for monomeric polyglutamine. Dots of different colors correspond to chains of different length. Representative points are marked using stars and the corresponding structures are shown in cartoon representation. Graphics were generated using VMD.[36] Note 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.

We can therefore conclude that at high values of f_β the restraints on f_β provide a diverse ensemble of structures. We can also conclude that the most prominent conformations high in β-content for both metrics are almost all β-sheets or β-hairpins, but never β-helices. For the latter to be stable, the peptides may have to be longer.[39] Conversely, conformations with high f_β-values but low

DSSP E-scores appear completely disordered. The latter are high in β-content only when measured by ϕ,ψ-propensities but lack the characteristic backbone-backbone hydrogen bonds.

Consequently, we asked if disordered structures contain large numbers of hydrogen bonds which are either unsatisfied or are assumed to be satisfied by the solvent? In Figure 5.9, we plot mean hydrogen bond numbers corresponding to the backbone and sidechain acceptor oxygen atoms, respectively:

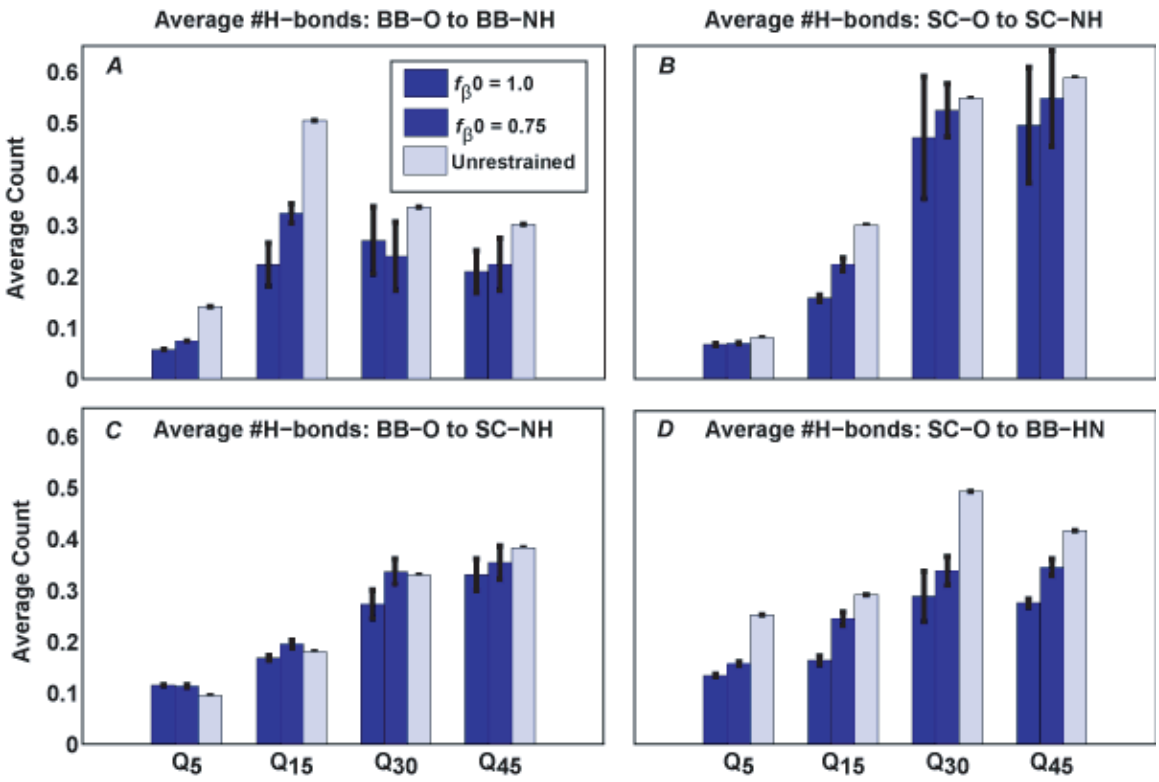

Figure 5.9: Hydrogen bond statistics around acceptor atoms for monomeric polyglutamine. Data are shown for hydrogen bonds around the backbone (Panels A and C) and sidechain (Panels B and D) acceptor oxygen atoms, respectively. BB denotes backbone and SC denotes sidechains; Values are shown for T=298K 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 consistent with the definition of the DSSP E-scores.[35]

The data are parsed into backbone-backbone, sidechain-backbone, and sidechain-sidechain terms. They reveal three main trends. First, all four terms contribute significantly and roughly equally to the total hydrogen bond registry. In agreement with the large spread observed in Figure 5.8 along the DSSP-axis, no obvious preference for specific intramolecular contacts is seen even in the presence of restraints. Second, summing up around the acceptor sites, the mean occupancy is typically less than unity indicating that for all chains studied here we assume a substantial fraction of hydrogen bonds are satisfied by the (implicit) solvent. This is simply saying that the interface with solvent is non-negligible. The surface-to-volume ratio (R_{SV}) decreases with increasing chain length and for longer chains more peptide donor atoms are found around the acceptor atoms on average. This explains the general increase in values with increasing N, including the overall very low values for Q_5, which is simply too short a peptide to form intramolecular hydrogen bonds. Third, restraints to high values of f_β seem to increase the number of solvent-exposed hydrogen bond acceptors indicated by the generally lower values. This is consistent with the observed swelling of chains, in particular for $f_\beta^0 = 1.0$ (see Figure 5.10 in V.4.5). Even though the backbone torsions are forced into the β-basin, the peptides prefer to take advantage of all possible intramolecular contacts. This is of course not unexpected; specific sidechain-interactions were part of Perutz's earliest models

for aggregates of polyglutamine and are found prominently in microcrystals of short amyloidogenic peptides.[8,40]

V.4.5. Coil-to-Globule Transition

From previous work (see Chapters II and IV)[1,13,37] we know that polyglutamine exists as compact, disordered globules under poor solvent conditions and that the chains undergo a cooperative swelling transition with increasing solvent quality which can be modulated by temperature. The data in Figure 5.9 provide evidence that in the presence of conformational restraints on f_β the solute-solvent interface is significantly increased. It is therefore worth asking whether the nature of the coil-to-globule transition is significantly altered or whether the transition is entirely obliterated by the presence of restraints.

Figure 5.10 plots the normalized radius of gyration as a function of simulation temperature. Two trends are apparent in the data shown in Figure 5.10. First, the coil-to-globule transition 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 $f_\beta \geq 0.75$ but significantly smaller if $f_\beta \leq 0.25$. Both the increased cooperativity and the divergence of the baselines become more pronounced with increasing chain length. In particular,

the lack of swelling due to restraints in the collapse regime is observed only for Q_{45}, but not for Q_{15} and Q_{30}.

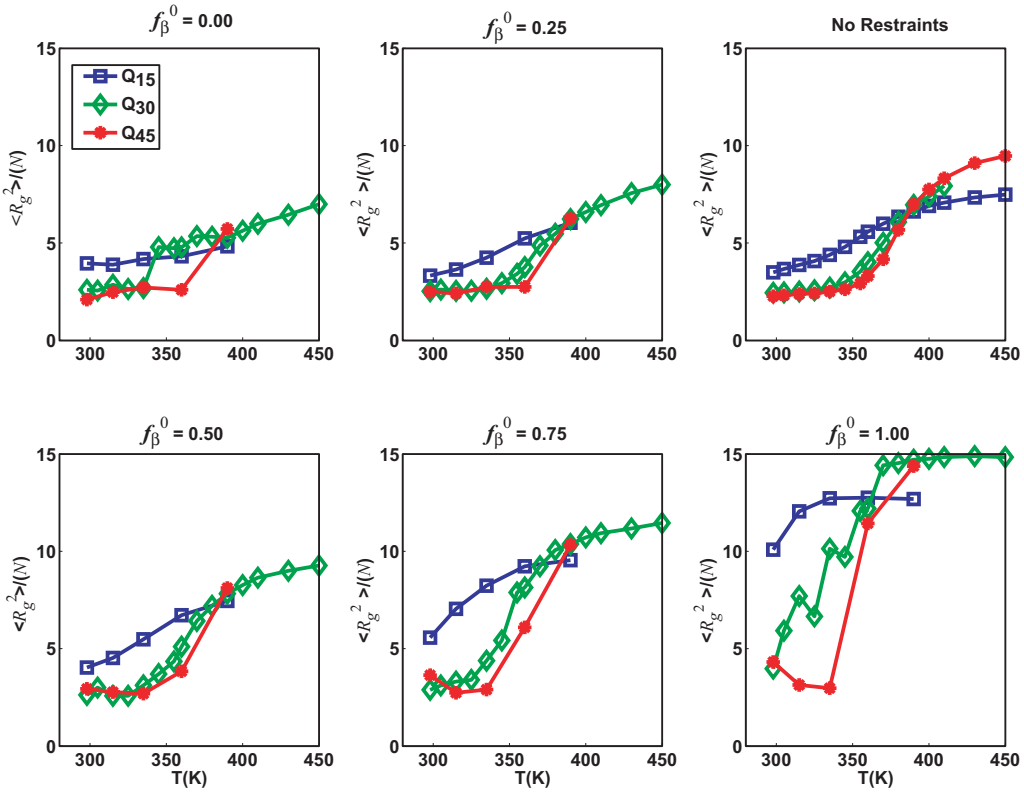

Figure 5.10: **Coil-to-globule transitions for monomeric polyglutamine in the presence of restraints on f_β.** Data are shown as a function of simulation temperature and for several different values of f_β^0 indicated above each panel. Data include the unrestrained case for which the most data are available. Three different chain lengths are shown: Q_{15}, Q_{30}, and Q_{45}. Q_5 was omitted since the peptide is too short to undergo a distinct swelling transition.

The above discussion is consistent with the result in Figure 5.9 which shows the most pronounced depletion of hydrogen bonds with restraint potentials for a shorter chain, *viz.* Q_{15}. The results in Figure 5.10 may be understood as follows: when f_β^0 is high and *local* order is increased, the effective stiffness of the

chain increases due to the enthalpic cost associated with leaving the β-basin. The collapse transition is altered: in the absence of restraints, dense packing of a highly flexible coil occurs, whereas high values of f_β^0 lead to a scenario in which semi-rigid rods are assembled into a compact body in a topologically frustrated manner. The observed increase in cooperativity is consistent with predictions from the polymer literature on an analogous model system.[41]

V.4.6. Dimerization Propensity in the Presence of β-Bias

Next, we turn our attention to the question of the dimerization propensity of polyglutamine. Previously we found that – at concentrations approaching those of typical *in vitro* experiments – dimerization of polyglutamine occurs spontaneously at room temperature (see Figure 4.4).[1] We showed that within the spontaneous fluctuations accessible to the system there is no structural signature associated with this homotypic association to occur. In other words, disorder is maintained from the monomer to the dimer, and this disorder includes the interface. Here, we first focus on the spontaneity of dimerization when high levels of β-content are enforced through the application of restraints. We consequently simulated dimers of polyglutamine of the same four lengths as before at varying temperatures and values for the target restraint value, f_β^0, specifically f_β^0=0.75 and f_β^0=1.0. Now, the restraint acts on the net β-content of the system, *i.e.*, it inherently averages over the two molecules in the system.

If the homogeneous nucleation model[18,19] were correct, one would predict that associations involving ordered species are extremely favorable and that the

ordered nucleus is capable of preferentially biasing a bound monomer toward a similarly ordered state. We re-employ the excess interaction coefficient B_{22}, which resembles a normalized second virial coefficient (see Equation 4-3) with the slightly different value for the θ-point obtained here: $T_\theta \approx 390$K (see Figure 5.10). As was detailed in Chapter IV, there we used slightly modified LJ parameters and a reduced partial charge set which alters the results somewhat (compare Figures 4.1 and 5.10 for the unrestrained case). Figure 5.11 shows B_{22} as a function of chain length and temperature and in the presence or absence of restraints on f_β:

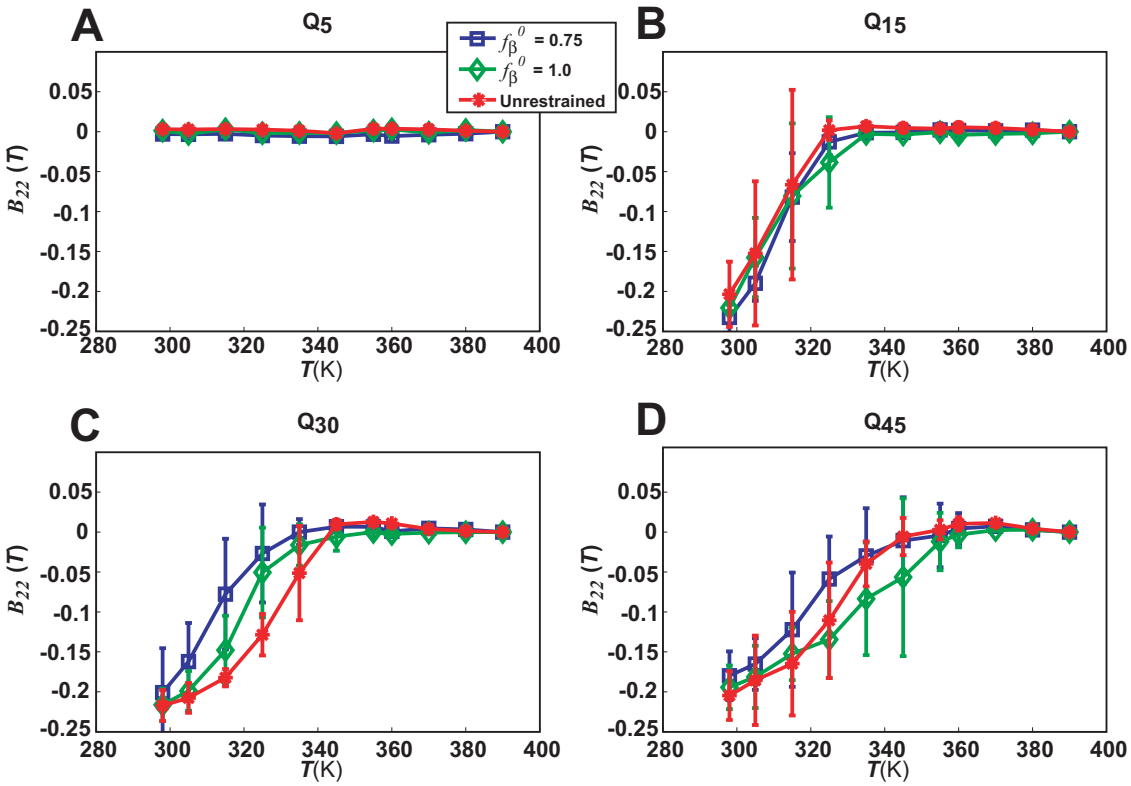

Figure 5.11: The excess interaction coefficient in the presence of restraints on f_β. Each panel shows B_{22} as a function of temperature extracted from simulations with unrestrained chains and simulations where each chain in the simulation has a target restraint of f_β^0=0.75 or f_β^0=1.0.

In general we find no evidence of β-rich structures promoting the association of polyglutamine. Panel A reveals that Q_5 at the simulated concentration is non-associative. Figure 5.7 showed that there is almost no penalty to access β-rich states yet dimerization is never spontaneous. This result questions the relevance of work on short peptides extracted from amyloidogenic sequences for obtaining insight about the aggregation pathways of longer peptides.[42,43] For Q_{15}, as shown in Panel B of Figure 5.11, a comparison of data using unrestrained chains to data from simulations with restraints of f_β^0=0.75 and f_β^0=1.0 indicates no significant differences. At the lowest temperatures, B_{22} is negative for all three cases. It increases monotonically with increasing temperature and reaches zero at about 340K indicating a complete elimination of intermolecular associations from this temperature onward.

As in previous work,[1] we conclude that dimer formation is only possible for a long enough chain in a poor enough solvent, *i.e.*, at a low enough temperature. From Figure 5.10, we note that the coil-to-globule transition as a function of decreasing temperature is maintained with similar transition temperatures even in the presence of conformational restraints but that – at least for Q_{15} – the interface with the surrounding solvent is increased in the collapse regime (see Figure 5.9). This, however, does not translate into increased associativity with respect to the unrestrained chains as Figure 5.11 makes clear. Instead, the actual poorness of the solvent, *i.e.*, the simulation temperature remains the main determinant of the spontaneity of intermolecular associations.

Panels B and C of Figure 5.11 show the same sets of data for Q_{30} and for Q_{45}. In agreement with our previous results, longer chain lengths exhibit a shift of the transition region for eliminating association to higher temperatures. Within the statistical accuracy of the data, chains with restraints appear to be indistinguishable from the unrestrained ones. For Q_{30}, the restraints might actually lead to a diminution of dimer formation compared to the unrestrained, disordered globules, especially at f_β^0=0.75. Taken together, these data indicate very clearly that the association propensity of polyglutamine peptides over a relevant range of chain lengths is not strongly tied to any structural motifs or even to the size of the solute-solvent interfaces. They therefore provide no supporting evidence for a structure-centric, homogeneously nucleated aggregation mechanism. Instead, we expect to see heterogeneity at the level of small oligomers which would show little to no preference for the secondary structure content or even polymeric state of the recruited monomers. This observation is ultimately in line with the extremely low solubility of polypeptides devoid of charged residues in a wide variety of aqueous solution conditions.

V.4.7. Intermolecular Interfaces in the Presence of β-Bias

Finally, we interrogated the nature of the associations between polyglutamine chains in the presence of conformational restraints on both chains. Specifically, we asked if the insensitivity of B_{22} to the presence or absence of restraints can be explained simply as an invariance of the interface size. Based on Figures 5.9 and 5.10 we suggested that monomers possess increased interfaces with the surrounding solvent at high values of f_β^0. To illustrate this point,

we linearly fitted the system energy (excluding the restraint potential) against chain length according to:

$$\frac{\left\langle U_{total} - U_{f_\beta} \right\rangle}{N} = C_1\left(T, f_\beta^0\right) + C_2\left(T, f_\beta^0\right) \cdot N^{-1/3} \qquad (5\text{-}4)$$

Here, N is chain length, U_{total} is the total system energy, $U_{f\beta}$ is the restraint potential energy, and C_1 and C_2 are the volumetric and surface terms, respectively. Figure 5.12 shows plots of C_1 and C_2 as a function of temperature for different values of f_β^0:

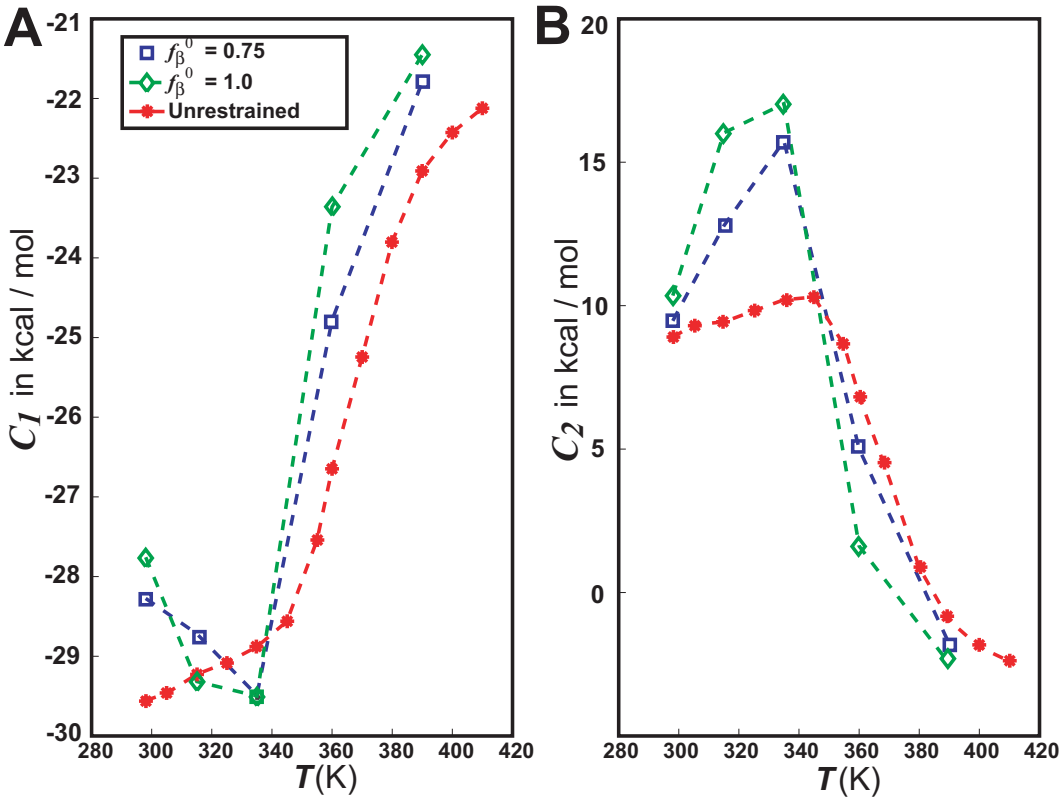

Figure 5.12: Energy density C_1 (Panel A) **and surface energy term C_2** (Panel B) **for monomeric polyglutamine**. Data were obtained for unrestrained polyglutamine and two other simulations with restraints on f_β. The quality of the fits underlying these data cannot be accurately assessed since they are fits to data from only three chain lengths (we

have to exclude Q_5 excluded from analysis since it is too short for a volumetric term to contribute). Instead the robustness of the analysis is tested through its ability to capture well-defined limits (see text).

When compared to the unrestrained simulations, it becomes clear that in the collapse regime the energy densities C_1 are generally less favorable in the presence of restraints indicating the high free energy of structures with high values of f_β (see Figure 5.7). Additionally, for a given temperature the values for C_2 are generally more positive indicating a less favorable, *i.e.*, simply a larger interface with the surrounding solvent.

For higher temperatures two observations are noteworthy. First, as $T \rightarrow T_\theta \approx 390$K, the value of C_2 reliably approaches zero, independent of whether restraints are applied or not. A value of zero for the surface energy indicates that the interface has become indifferent with respect to interactions with either solvent or the chain. This is nothing but the Flory-definition of the θ-state and hence a satisfying if expected result. Second, for temperatures beyond T_θ, C_1 appears to converge toward a value of roughly -20kcal/mol. This is the approximate mean-field energy for a fully solvated glutamine residue within the ABSINTH Hamiltonian (see Table 3.a), which constitutes the expected limiting value for a chain preferring chain-solvent interactions to chain-chain interactions. These two results indicate the robustness of the analysis. We conclude that in the presence of restraints toward high values of f_β the chains do indeed swell and form a less favorable interface with the surrounding solvent. However, these two features do not translate into differences in B_{22} vis-à-vis the unrestrained chains.

Instead, they may contribute to conformational rearrangements, evidence for which is presented next.

To show that the dimer interface at high values of f_β promotes β-content, Figure 5.13 shows bar plots of the mean DSSP E-scores observed in the dimer simulations along with those encountered in the monomer simulations:

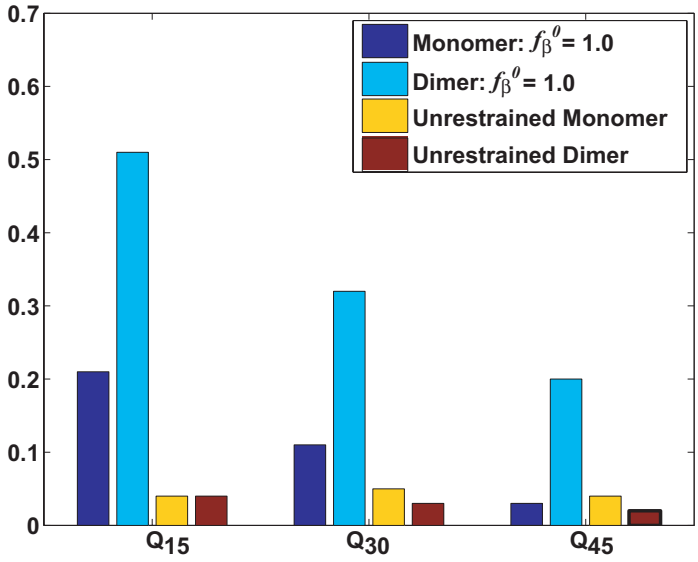

Figure 5.13: Bar plots comparing the average fractional DSSP-E scores between monomer and dimer simulations. Data are shown for 298K and for three chain lengths using data from simulations with unrestrained chains as well as data from simulations for f_β^0=1.0.

The second molecule in the simulation system significantly increases the prevalence of canonically hydrogen-bonded, β-rich structures when compared to the monomer case. This difference seen upon introducing a second molecule into the simulation system is strongly suggestive of the intermolecular interface being responsible for promoting β-sheet content. Consistent with Figure 5.8,

values for the shorter chains are generally higher. Figure 5.14 is analogous to Figure 5.8 and uses data from dimer simulations:

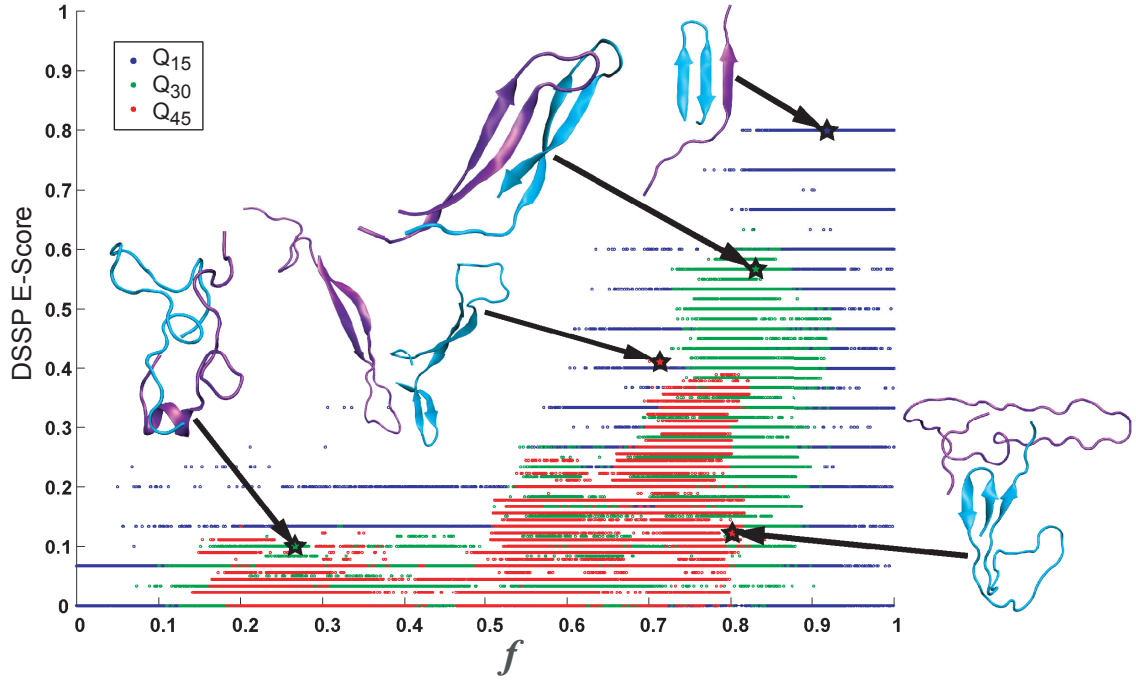

Figure 5.14: Scatter plot of DSSP E-scores and f_β for dimers of polyglutamine. This plot is analogous to Figure 5.8. Note that we only performed dimer simulations for three different conditions: two restrained sets for specific values of f_β^0 and the unrestrained case. Therefore, the scatter plot here is much less homogeneously populated compared to the one in Figure 5.8. Data are obtained at 298K.

Comparisons of example structures shown in Figures 5.8 and 5.14 along with the data in Figure 5.13 illustrate that dimerization in a poor solvent promotes the formation of canonical β-secondary structure providing the monomer is predisposed to high β-content. Non-specific collapse, however, imposes a requirement for bends and turns on the polypeptide backbone and thereby

precludes efficient and extensive formation of β-hydrogen bonds. Both f_β and DSSP E-scores are low in the absence of structural restraints.

To further illustrate the ability of *inter*molecular contacts to promote β-content, Figure 5.15 shows a bar plot similar to Figure 5.9 for *inter*molecular hydrogen bonds:

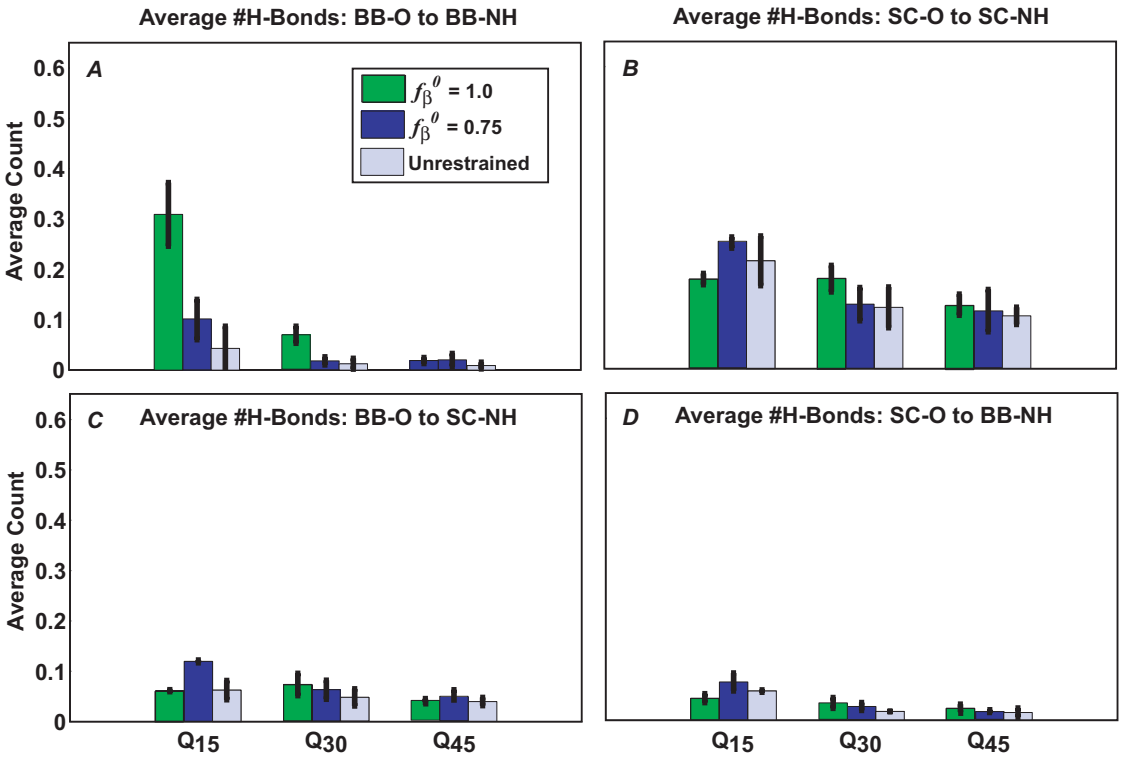

Figure 5.15: Average number of *inter*molecular hydrogen bonds per acceptor oxygen atoms. This plot is similar to Figure 5.9 except that the hydrogen bond statistics are shown for simulations with two molecules. Only intermolecular hydrogen bonds are shown. Q_5 is excluded from this plot since no intermolecular hydrogen bonds are detected in these simulations. Data are obtained at 298K.

First, it is clear that intermolecular hydrogen bonds are formed predominantly by the glutamine sidechains independent of restraints. This is

intuitive as the sidechains are generally closer to the surface of the globule. Second, there is significant enhancement in intermolecular backbone-backbone hydrogen bonds for Q_{15} and Q_{30} when $f_\beta^0 = 1.0$. This result is entirely congruent with the increased DSSP E-Scores (see Figure 5.13). Nonetheless, some amount of disorder is maintained even under these conditions as indicated by the non-negligible contributions from all possible intermolecular hydrogen bonds.

As a final point, we wish to emphasize that the increased prevalence of canonical β-sheets does not coincide with a significant reduction in the free energy penalty to access states with large f_β. At 298K, the per-molecule PMF obtained via WHAM for two Q_{30} molecules in the simulation system (data not shown) is virtually identical to the one obtained for the monomer case (see Figure 5.7). We therefore conjecture that it might take much larger oligomers to provide an environment in which β-rich conformations are spontaneously accessed. This would presumably happen on the *inside* of such larger oligomers.

V.5. Summary and Discussion of a Putative Role of β-Secondary Structure in Polyglutamine Aggregation

We have shown that polyglutamine peptides in a length range encompassing the threshold length for Huntington's disease do not easily adopt conformations rich in β-content, *i.e.*, conformations that have been proposed as putative aggregation nuclei. Our estimate for the free energy barrier for nucleation is roughly consistent with the literature estimate,[19] but shows an opposite trend with chain length. We observe that longer chains are less likely to

adopt conformations rich in canonical β-secondary structure and therefore dispute the validity of explaining the age-of-onset dependence on chain length with that free energy barrier. We find that structures which are forced into states characterized by partial swelling, larger solute-solvent interfaces, and increased β-content do not specifically promote aggregation. While structural differences are obvious, the peptides seem to associate to similar extents independent of those differences. This observation points to the poorness of the solvent as the invariant driving force for promoting aggregation of these homopolymers – a result entirely consistent with Chapter IV. Even though high β-content remains thermodynamically as unfavorable at the dimer level as at the monomer level, we do find that intermolecular interfaces appear to promote the formation of canonical, backbone-driven β-secondary structure as it is found in amyloid fibrils.

From our data, we infer that the role of β-secondary structure for the aggregation of homopolymeric polyglutamine is not so much that of the structural hallmark of an aggregation-competent species but rather that of the favored conformation in peptide-rich phases such as large oligomers and fibrillar aggregates. Figure 5.16 (compare Figure 4.12) outlines how the formation of insoluble, fibrillar aggregates could proceed through a stage characterized by intermediate, soluble aggregates composed of collapsed, disordered monomers and disordered interfaces between them:

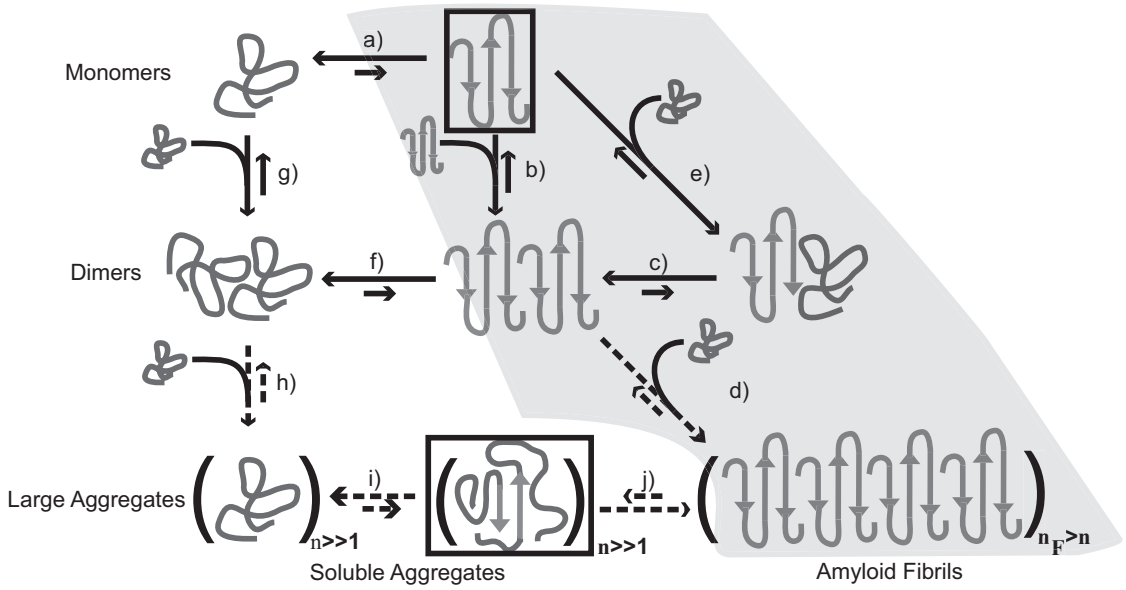

Figure 5.16: Schematic of possible aggregation pathways for polyglutamine *in vitro*. n denotes the number of polyglutamine molecules within a disordered aggregate and n_F denotes the number of polyglutamine molecules within an ordered amyloid fibril. The ordered amyloid fibril rich in β-sheets is shown 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.[18]

In the schematic shown in Figure 5.16 we quantified the thermodynamics of step (a), which indicates that the formation of ordered conformations is thermodynamically unfavorable. This is an extension of the work in Chapter IV which remained agnostic about the barrier heights and their chain length dependence implied in 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) – as shown in Figure 5.11.

However, the likelihood that chains will sample the associations shown in step (b) is very small because this is tied to the equilibria in steps (a) and (f) and requires the population of the conformations with high β-content (Figure 5.7). The aggregates achieved in step (h) are likely to be large (in terms of n) and exhibit spherical, "liquid-like", "molten", or micellar organization of polyglutamine chains around each other. Such a phenomenology was observed for other aggregation-prone polypeptides.[44-48] Step (i) depicts a slow conformational conversion of individual / small numbers of chains to β-sheets. Such a conversion is supported by the idea that the peptide aggregate represents a θ-solvent for an individual molecule, in which the swelling necessary for large intermolecular interfaces in extended β-sheets is greatly facilitated.[24] This slow conversion 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, precipitated amyloid fibril. Steps (a), (b), and (g) are anchored in the collection of data generated in this work and previous work (Chapters II and IV). However, the reversible associations depicted in step (h) and the conformational conversions depicted in step (i) are yet to be tested.

It has been argued, but is of course difficult to prove rigorously, that polyglutamine aggregation is much more complex than a simple homogeneous nucleation model would suggest.[22,25,27] This is particularly true if non-negligible sequence context is introduced. Much work has been carried out in the context of completely unrelated proteins such as myoglobin,[49] GST,[50] CRABP,[27] or thioredoxin.[17] Collectively, these studies suggest a generality to the behavior and

consequences of expanded polyglutamine regions. However, they also suggest that the details of the aggregation mechanism might depend very strongly on sequence context. So far, the work in this thesis has sought to isolate the *intrinsic* preferences of polyglutamine. The results have led us to the mechanistic view sketched in Figure 5.16. The next step is concerned with how and by how much those intrinsic preferences are modulated by the presence of both N- and C-terminal flanking sequences as they might occur in proteolytic fragments found natively *in vivo*. Part of this work is presented in Chapter VI.

V.6. Bibliography

1. Vitalis, A.; Wang, X.; Pappu, R. V. J Mol Biol 2008, 384(1), 279-297.

2. Vitalis, A.; Lyle, N.; Pappu, R. V. Biophys J 2009, *in press*.

3. Scherzinger, E.; Lurz, R.; Turmaine, M.; Mangiarini, L.; Hollenbach, B.; Hasenbank, R.; Bates, G. P.; Davies, S. W.; Lehrach, H.; Wanker, E. E. Cell 1997, 90(3), 549-558.

4. Chen, S. M.; Berthelier, V.; Hamilton, J. B.; O'Nuallain, B.; Wetzel, R. Biochemistry 2002, 41(23), 7391-7399.

5. Perutz, M. F.; Finch, J. T.; Berriman, J.; Lesk, A. Proc Natl Acad Sci U S A 2002, 99(8), 5591-5595.

6. Sikorski, P.; Atkins, E. Biomacromolecules 2005, 6(1), 425-432.

7. Nelson, R.; Sawaya, M. R.; Balbirnie, M.; Madsen, A. O.; Riekel, C.; Grothe, R.; Eisenberg, D. Nature 2005, 435(7043), 773-778.

8. Sawaya, M. R.; Sambashivan, S.; Nelson, R.; Ivanova, M. I.; Sievers, S. A.; Apostol, M. I.; Thompson, M. J.; Balbirnie, M.; Wiltzius, J. J. W.; McFarlane, H. T.; Madsen, A. Ã.; Riekel, C.; Eisenberg, D. Nature 2007, 447(7143), 453-457.

9. Diaz-Avalos, R.; Long, C.; Fontano, E.; Balbirnie, M.; Grothe, R.; Eisenberg, D.; Caspar, D. L. D. J Mol Biol 2003, 330(5), 1165-1175.

10. Paravastu, A. K.; Leapman, R. D.; Yau, W. M.; Tycko, R. Proc Natl Acad Sci U S A 2008, 105(47), 18349-18354.

11. Bernstein, S. L.; Wyttenbach, T.; Baumketner, A.; Shea, J. E.; Bitan, G.; Teplow, D. B.; Bowers, M. T. J Am Chem Soc 2005, 127(7), 2075-2084.

12. Marina, G. B.; Kirkitadze, D.; Lomakin, A.; Vollers, S. S.; Benedek, G. B.; Teplow, D. B. Proc Natl Acad Sci U S A 2003, 100(1), 330-335.

13. Vitalis, A.; Wang, X.; Pappu, R. V. Biophys J 2007, 93(6), 1923-1937.

14. Marchut, A. J.; Hall, C. K. Prot Struct Funct Bioinf 2007, 66(1), 96-109.

15. Khurana, R.; Gillespie, J. R.; Talapatra, A.; Minert, L. J.; Ionescu-Zanetti, C.; Millett, I.; Fink, A. L. Biochemistry 2001, 40(12), 3525-3535.

16. Zurdo, J.; Guijarro, J. I.; Jiménez, J. L.; Saibil, H. R.; Dobson, C. M. J Mol Biol 2001, 311(2), 325-340.

17. Nagai, Y.; Inui, T.; Popiel, H. A.; Fujikake, N.; Hasegawa, K.; Urade, Y.; Goto, Y.; Naiki, H.; Toda, T. Nat Struct Mol Biol 2007, 14(4), 332-340.

18. Chen, S. M.; Ferrone, F. A.; Wetzel, R. Proc Natl Acad Sci U S A 2002, 99(18), 11884-11889.

19. Bhattacharyya, A. M.; Thakur, A. K.; Wetzel, R. Proc Natl Acad Sci U S A 2005, 102(43), 15400-15405.

20. Carrell, R. W.; Lomas, D. A. Lancet 1997, 350(9071), 134-138.

21. Uversky, V. N.; Fink, A. L. Biochim Biophys Acta, Prot Proteomics 2004, 1698(2), 131-153.

22. Lee, C. C.; Walters, R. H.; Murphy, R. M. Biochemistry 2007, 46(44), 12810-12820.

23. Akasaka, K.; Latif, A. R. A.; Nakamura, A.; Matsuo, K.; Tachibana, H.; Gekko, K. Biochemistry 2007, 46(37), 10444-10450.

24. Pappu, R. V.; Wang, X.; Vitalis, A.; Crick, S. L. Arch Biochem Biophys 2007, 469(1), 132-141.

25. Bernacki, J. P.; Murphy, R. M. Biophys J 2009, 96(7), 2871-2887.

26. Bulone, D.; Masino, L.; Thomas, D. J.; San Biagio, P. L.; Pastore, A. PLoS ONE 2006, 1(1), e111.

27. Ignatova, Z.; Thakur, A. K.; Wetzel, R.; Gierasch, L. M. J Biol Chem 2007, 282(50), 36736-36743.

28. Takahashi, T.; Kikuchi, S.; Katada, S.; Nagai, Y.; Nishizawa, M.; Onodera, O. Hum Mol Genet 2008, 17(3), 345-356.

29. Temussi, P. A.; Masino, L.; Pastore, A. EMBO J 2003, 22(3), 355-361.

30. Tran, H. T.; Mao, A.; Pappu, R. V. J Am Chem Soc 2008, 130(23), 7380-7392.

31. Roux, B. Comput Phys Commun 1995, 91(1-3), 275-282.

32. Shankar Kumar, J. M. R., Djamal Bouzida, Robert H. Swendsen, Peter A. Kollman. J Comput Chem 1992, 13(8), 1011-1021.

33. Sugita, Y.; Kitao, A.; Okamoto, Y. J Chem Phys 2000, 113(15), 6042-6051.

34. Hobohm, U.; Scharf, M.; Schneider, R.; Sander, C. Protein Sci 1992, 1(3), 409-417.

35. Kabsch, W.; Sander, C. Biopolymers - Peptide Science Section 1983, 22(12), 2577-2637.

36. Humphrey, W.; Dalke, A.; Schulten, K. J Mol Graph 1996, 14(1), 33-38.

37. Crick, S. L.; Jayaraman, M.; Frieden, C.; Wetzel, R.; Pappu, R. V. Proc Natl Acad Sci U S A 2006, 103(45), 16764-16769.

38. Walker, F. O. Lancet 2007, 369(9557), 218-228.

39. Ogawa, H.; Nakano, M.; Watanabe, H.; Starikov, E. B.; Rothstein, S. M.; Tanaka, S. Comput Biol Chem 2008, 32(2), 102-110.

40. Perutz, M. F. Mol Med 1995, 1(7), 718-721.

41. Sear, R. P. J Chem Phys 1997, 107(18), 7477-7482.

42. Santini, S.; Mousseau, N.; Derreumaux, P. J Am Chem Soc 2004, 126(37), 11509-11516.

43. Tartaglia, G. G.; Cavalli, A.; Pellarin, R.; Caflisch, A. Protein Sci 2005, 14(10), 2723-2734.

44. Lomakin, A.; Asherie, N.; Benedek, G. B. AIP Conf Proc 2003, no.690, 390-391.

45. Krishnan, R.; Lindquist, S. L. Nature 2005, 435(7043), 765-772.

46. Lomakin, A.; Chung, D. S.; Benedek, G. B.; Kirschner, D. A.; Teplow, D. B. Proc Natl Acad Sci U S A 1996, 93(3), 1125-1129.

47. Kumar, S.; Udgaonkar, J. B. J Mol Biol 2009, 385(4), 1266-1276.

48. Ceru, S.; Zerovnik, E. FEBS Lett 2008, 582(2), 203-209.

49. Tanaka, M.; Morishima, I.; Akagi, T.; Hashikawa, T.; Nukina, N. J Biol Chem 2001, 276(48), 45470-45475.

50. Masino, L.; Kelly, G.; Leonard, K.; Trottier, Y.; Pastore, A. FEBS Lett 2002, 513(2-3), 267-272.

CHAPTER VI. SEQUENCE CONTEXT DEPENDENCIES IN POLYGLUTAMINE AGGREGATION: AN ILLUSTRATION USING THE N-TERMINUS OF HUNTINGTIN

VI.1. Preamble

Our interest in the sequence context dependencies of polyglutamine aggregation was initially triggered by *in vitro* studies on oligoproline. Two studies showed that a C-terminal attachment of an oligoproline segment to polyglutamine could alter the aggregation rates, the morphology of resultant aggregates, and even the conformational ensembles of soluble peptides.[1,2] In the host protein of HD, huntingtin, the polyQ-expansion is C-terminally flanked by a proline-rich segment starting with eleven consecutive proline residues (see Table 1.c). This suggests a direct relevance of these studies to the pathogenesis of HD *in vivo*. As is detailed in VI.2, it is not intuitively clear that such relevance can be upheld due to the diversity and length of fragments encountered *in vivo*, and their strong differential effects on pathogenesis when studied individually.

Nonetheless, these and similar *in vitro* data[1-3] appeared to present a worthwhile challenge to our ability to realistically model the properties of polyglutamine in the presence of flanking sequences. Work in Chapters IV[4] and V[5] had exclusively focused on the intrinsic properties of polyglutamine but provided no further assessment of the ability of the ABSINTH implicit solvation model presented in Chapter III[6] to realistically describe the properties of more complex peptides. In early 2008, Tim E. Williamson – at that time a graduate

student in the laboratory – began simulation work on sequence constructs of the type Acetyl-$Q_{15}P_{11}$-N-methylamide. A detailed account is not given here but may – upon completion – appear elsewhere in the future. Briefly, studies at the monomer level revealed that these peptides continue to undergo a well-defined coil-to-globule transition as a function of simulation temperature (see Chapter IV) and that they remain disordered. Analysis of the proline segment yielded that, with the ABSINTH force field and a Monte Carlo sampling approach, large fractions of the amide bonds of the proline residues sampled the *cis*-conformation and the oligoproline stretch overall appeared surprisingly flexible. We suspected this to be a simulation artifact based on the (albeit highly variable) estimates for the *cis*-content of proline-based peptides extracted from various experimental techniques. We revised our modeling of proline slightly, and subsequent results exhibited significantly lower *cis*-contents for proline-based peptides. However, a satisfactory, global agreement with data on model peptides extracted from NMR spectroscopy[7] was not obtained. The peptides still appeared to be too flexible and to populate the *cis*-configuration in excess of experimentally derived expectation. This coincided with a developing interest in the N-terminal flanking sequence encountered in huntingtin and the project was temporarily suspended.

In early 2009, Emma Morrison – at that time a rotation student – attempted to quantify our ability to correctly model proline using a different set of model peptides and a different set of experimental data.[8,9] In direct contradiction of the previous result, she found that the simulated populations of the *cis*-

configuration of the amide bond were generally too low. This point presented an impasse: MD-based *in silico* approaches with putatively more accurate force fields are rendered infeasible due to the extremely long timescale needed for *cis*/*trans*-isomerization. Experimental data seem to be mutually inconsistent with each other and/or depend crucially on the detailed nature of the model peptide employed. Guidance is hence limited in extracting physical principles to inform an improved model for proline. Moreover, a strong case has been made that electronic effects[10] are indispensable in explaining the intricate nature of proline, which we would only be able to capture empirically. Given these observations, we decided that it would remain incredibly difficult to distinguish simulation artifact from relevant result in studies of proline-rich systems; hence, all our focus shifted to the N-terminal segment.

This chapter is based entirely on simulations run and data analyzed by Tim E. Williamson. His contributions are as follows: he established a simulation protocol for obtaining reliable data on peptides of the type Nt17-Q_N. Partially based on suggestions and continued discussions he analyzed the data in a manner reflecting the heteropolymeric nature of the peptides. With the exception of Figures 6.1 and 6.4, all the figures in this chapter were created by him. The text in VI.3 and VI.4 is partially based on a skeletal manuscript prepared by him.[11] Scott L. Crick, a graduate student in the laboratory, performed the CD spectroscopy measurements presented in VI.4.1 and Figure 6.4. Peptides were obtained as a generous gift from Trevor P. Creamer at the University of Kentucky, Lexington (see VI.3.3).

VI.2. Introduction to Sequence Context Dependencies in HD

We touched upon the issue of proteolysis in Chapter I, specifically in I.2.4 and I.2.7. Almost all *in vivo* studies report a differential pathogenic effect in animal or cellular models of exonic CAG repeat diseases if different sequence constructs are expressed.[12-17] Even though there may be exceptions, it is commonly thought that the full-length, mutant host protein (see Table 1.b in I.2.3) is not the dominant toxic species. Instead, it appears as if proteolytic fragments derived from the expanded disease protein are more likely disease causing agents ("toxic fragment" hypothesis).[13,18,19] Following I.2.4, this may be understood conceptually: the presence of the folded domain solubilizes the mutant protein and partially protects the PQCS from the ill effects which would be induced if the polyQ-expansion were to be cleaved off. A mechanism for this would be similar to what applies to all folded proteins: hydrophobic residues for instance are tolerated if they are (at least partially) sequestered away and hence not available for potentially pathological interactions. Furthermore, the hydrophilic surface of globular proteins will prevent molecular chaperones from interacting with sequestered hydrophobic stretches and stress on the PQCS is minimal. But what happens if a homopolymeric, intrinsically insoluble, disordered, and aggregation-prone expansion is introduced? Clearly, the resultant properties will depend on the relative sizes of the protective host sequence and the mutant expansion: such dependence is demonstrated in VI.4.

Table 1.c (see I.2.7) shows that the sequences surrounding the polyQ-expansion in the host proteins for the nine exonic CAG repeat diseases share no

similarity except that they exhibit some prevalence for low complexity sequences. Most experimental research has been performed on HD, and hence it becomes our focus here: huntingtin (htt) is a large protein of roughly 350kD and mostly unknown function (see Table 1.b). The first exon on the gene (exon1) contains the CAG repeat. Short N-terminal sequence constructs encompassing exon1 have been used to demonstrate several specific protein-protein interactions in *Drosophila*.[20] The identified interaction partners were confirmed to be genetic modifiers of the disease phenotype. N-terminal sequence constructs – unlike full-length htt – of varying length were shown to mediate the association with mitochondria; a pathogenic mechanism related to axonal transport and Ca^{2+} homeostasis was suggested.[21,22] It was demonstrated that the presence of the 17 residues N-terminal to the polyQ-expansion (Nt17: MATLEKLMKAFESLKSF) redirected polyQ-expanded fragments from their normal nuclear location to mitochondria and that it altered the peptides' *in vivo* aggregation behavior.[21] Consistent with the idea of Nt17 acting as a cytosolic retention signal, it was shown that SUMOylation of lysine residues located within the Nt17 stretch promotes both nuclear localization and subsequent interference with transcriptional regulation when compared to non-SUMOylated substrates.[23]

The above results strongly suggest that the impact of flanking sequences may be profound and that an exclusive focus on the intrinsic properties of polyglutamine may fail in explaining HD pathogenesis. It is however not easily possible to quantify and characterize *in vivo* fragments as they might occur during proteolysis of mutant htt. Hence, the relevance of results obtained with

artificial constructs may be questioned: proteolytic analysis of htt has in fact revealed that despite the presence of well-defined cleavage sites the resultant fragments form a heterogeneous population of peptides of differing lengths.[24,25] As a somewhat parsimonious guess, we nonetheless attempt to quantify the effects of the Nt17-fragment on the intrinsic properties of polyglutamine in this chapter. It is very likely to be present in actual disease-relevant constructs due to its immediate proximity to the polyQ-expansion which we still assume plays the dominant pathogenic role.

Thakur *et al*.[26] recently demonstrated that the previous results obtained for peptide constructs of the type $K_2Q_NK_2$[27,28] are strongly altered for peptides of the type Nt17-Q_NK_2: distinct spherical oligomers are seen in electron micrographs early, the aggregation overall proceeds faster, and the apparent nucleus size for homogeneous nucleation yields the nonsensical result of -1 indicating that homogeneous nucleation does not apply. Mutational analysis appears to identify a prominent role for hydrophobic residues in the Nt17-fragment in mediating the increased aggregation rate. NMR and CD data indicate the Nt17-fragment to be mostly disordered, although α-helical conformations are readily populated in 10% (v/v) TFE. Recent *in silico* work[29] suggests more prominent α-helicity. Figure 6.1 shows that the Nt17-fragment may form an amphipathic helix as suggested.[29] The structural models reveal that the presence of α-helical hydrogen bonds requires the two interfaces characterized by either hydrophobic groups or by charged groups to be slightly twisted. Nonetheless, the amphipathic character is clearly visible in either panel.

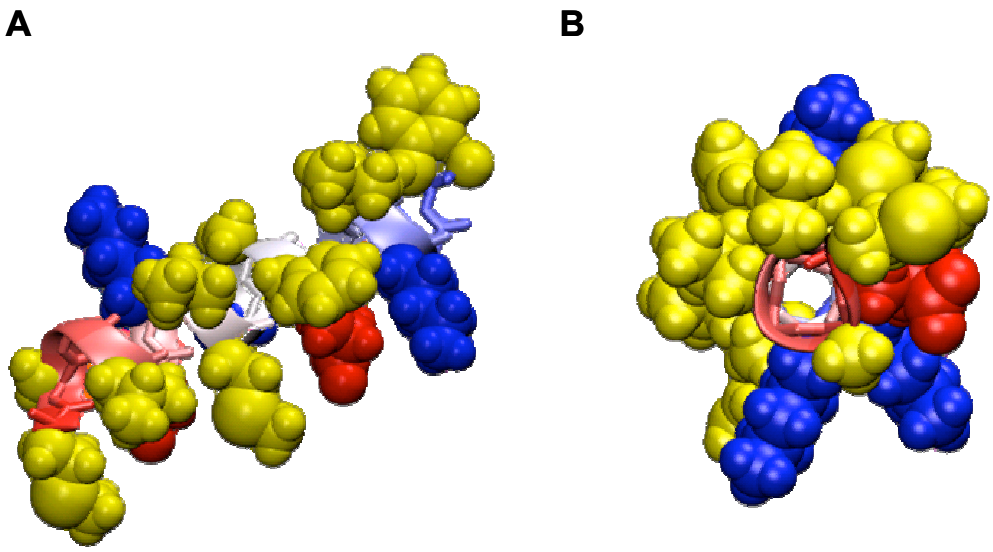

Figure 6.1: Renderings of the Nt17-peptide in idealized α-helical conformation. The structure was obtained by setting all backbone ϕ,ψ-angles to values of -57.8° and -47.0°, respectively. Panel A shows a side view and Panel B a top view analogous to a helical wheel diagram. Space-filling representations are chosen for select atoms: hydrophobic sidechains in yellow, negatively charged sidechains in red, and positively charged sidechains in blue. Graphics were generated using VMD.[30]

Here, we attempt to address the following questions pertaining to the modulation of polyglutamine's intrinsic preferences by the presence of the Nt17-fragment, *i.e.*, we investigate peptide constructs of the type Nt17-Q_N:

- Does the known secondary structure preference of the Nt17-fragment induce secondary structure in the polyQ-segment or does the latter remain disordered? Are any induced conformational biases dependent on N? Conversely, does the presence of polyQ-expansions of sufficient length alter the conformational ensembles populated by the Nt17-segment as suggested by Thakur *et al.*?[26]

- What types of intramolecular interfaces are formed? Are the hydrophobic residues effectively sequestered from solvent in constructs with large enough N or do they remain surface-exposed?
- Does polyglutamine spontaneously dimerize in the presence of the Nt17-fragment? What are the intermolecular interfaces formed? Is the association driven by hydrophobic contacts, by the polyQ-expansion, or by both?
- Is there a *differential* effect on associativity as observed by Thakur *et al*.?[26] Do glutamine homopolymers associate more or less readily and what is the impact of the flanking sequences used in *in vitro* experiments?

The rest of Chapter VI is structured as follows: first, we briefly introduce the simulation details which are largely similar to previous work (Chapters IV[4] and V[5]) and provide details for the CD spectroscopy experiments (VI.3). We then give a detailed account of our results (VI.4) and finish by discussing the impact of the results on *in vitro* and *in vivo* experiments, in particular the work of Thakur *et al*.[26] (VI.5).

VI.3. Simulation and Experimental Details

VI.3.1. System Setup and Conformational Sampling

The simulation methodology employed here agrees largely with what was presented in IV.3.1 and IV.3.2. Peptides were generally capped at the N- and C-termini. We studied three different types of constructs: acetyl-Nt17-Q_N-N-methylamide (Nt17 = METLEKLMKAFESLKSF), Acetyl-Q_N-N-methylamide, and Acetyl-$K_2Q_NK_2$-N-methylamide. The chains were modeled in atomic detail and

their interactions with each other and the solvent was described by the ABSINTH continuum solvation model[6] (see Chapter III). For all simulations, sampling was enhanced by the replica exchange (REX) method in temperature space.[31] The following temperature schedule was used: T = 298K, 305K, 315K, 325K, 335K, 345K, 355K, 365K, 375K, 385K. A summary of the simulations we performed is shown in Table 6.a:

Peptide	**Monomer**	**Dimer (100μM)**	**Dimer (500μM)**
Nt17	10	5	-
Nt17-Q_5	5	5	5
Nt17-Q_{10}	4	-	-
Nt17-Q_{15}	16	9	5
Nt17-Q_{20}	5	-	-
Nt17-Q_{25}	3	5	5
Nt17-Q_{30}	3	-	-
Nt17-Q_{35}	3	5	5
Nt17-Q_{45}	3	-	-
Q_5	-	3	-
Q_{15}	9	5	-
Q_{20}	3	-	-
Q_{25}	3	5	-
Q_{30}	3	-	-
Q_{35}	3	5	-
Q_{45}	3	-	-

Peptide	Monomer	Dimer (100μM)	Dimer (500μM)
K_2-Q_5-K_2	-	3	-
K_2-Q_{15}-K_2	-	3	-
K_2-Q_{25}-K_2	-	3	-
K_2-Q_{35}-K_2	-	5	-

Table 6.a: Extent of simulations studying the effects of the Nt17-fragment of huntingtin. For each peptide investigated, the table lists the number of independent sets of replicates obtained for a given condition. A single replicate is defined as a REX run encompassing ten individual replicas corresponding to ten different temperatures (see text). Each individual replica was run for $5.15 \cdot 10^7$ production steps after 10^7 steps of equilibration. These settings were needed to obtain quantitatively meaningful results for the longest peptides studied. Not all constructs were simulated under all conditions as indicated by dashes.

Markov chain Metropolis Monte Carlo (MC) simulations were performed in the canonical ensemble utilizing our in-house CAMPARI software.[32] The peptides were enclosed in a spherical droplet whose boundary was modeled as a stiff, one-sided harmonic potential according to Equation 4-1. At pH 7, the Nt17-Q_N constructs are expected to carry a net charge of approximately 1.0. To prevent strong simulation artifacts from electrostatic interactions modeled in an unrealistic milieu, we added explicit chloride counterions plus a number of sodium and chloride ions equal to a salt concentration of 2.5mM. To model the effects of peptide concentration we used two different droplet sizes in dimer simulations: 117.0Å and 200.0Å corresponding to peptide concentrations of 500 and 100μM,

respectively (see Table 6.a). All monomer simulations were performed in droplets of radius 200.0Å.

The presence of explicit counterions means that the sampling of rigid-body degrees of freedom in the MC methodology may be less efficient. We therefore employed the MC move set detailed in Table 6.b:

Move type	Parameters
Rigid-body	27% (50%, 10Å, 20°)
Random cluster	3% (50%, 10Å, 20°)
Omega (ω)	4.9% (90%, 5°)
Sidechain (χ_1, χ_2, χ_3)	21% (4x, 60%, 30°)
Backbone ϕ/ψ	44.1% (70%, 10°)

Table 6.b: Overview of the frequency of the different Monte Carlo moves sets used in simulations of polyQ-expanded Nt17. Random cluster moves attempt to perturb the rigid-body coordinates of two molecules simultaneously: two molecules are chosen at random, translated by a common vector, and rotated around their mutual center of mass. By preserving their relative orientation, tightly associated molecules can be sampled much more effectively. Please refer to the caption to Table 4.a for details about the other move types and their underlying parameters.

VI.3.2. Analysis of Simulation Data

Error Analysis

Snapshots of the system were saved every 5000 steps and intermolecular distances and polymeric properties were recorded every 500 steps. In general, errors were obtained by running simulations as multiple replicates (see Table

6.a). The standard deviations of computed ensemble averages from each trajectory yield standard errors. Those are typically reported in the figures in this chapter. The analysis in Figure 6.14 required rigorous error propagation of the standard errors through the linear regression analysis we performed.

Contact Analysis

We collected contact maps describing the frequency of observing an interaction between all residue pairs in Nt17-Q_N. Such data were collected every 500 simulation steps. A contact is counted between two residues only if those residues are not adjacent in sequence space and the minimum distance between any atom pair is less than 3.5Å. The residues of Nt17-Q_N are classified as being hydrophobic (M, A, F, L → "hydrophobes"), hydrophilic (K, E, S, T → "hydrophiles"), or glutamine. With increasing N, the probability of forming more contacts involving glutamine grows by default. We therefore defined a combinatorial prior assuming that the sets of residues are allowed to mix freely. There are N glutamine residues, eight hydrophobes, and nine hydrophiles in peptides of the type Nt17-Q_N. We corrected the combinatorial weights in the prior for the fact that next-neighbor contacts are excluded in the contact analysis. For example, there are (N+17)·(N+16)/2 - (N+16) total non-neighbor, intramolecular contacts. Of those, N·(N-1)/2 – (N-1) are glutamine-glutamine contacts. The ratio of the latter and former numbers defines the combinatorial prior that a given intramolecular contact is a glutamine-glutamine contact. Similarly, the remaining combinatorial weights can be obtained for all other possible contacts, both intra- and intermolecularly.

Clustering analysis

To identify clusters of structures within the trajectories of either one or two peptides of the type Nt17-Q_N, we utilized a variation of a bottom-up clustering algorithm[33] implemented in the GROMACS 4.0 utility *g_cluster*.[34] Structural clustering allows us to identify the most common conformations in the ensemble. A matrix of pairwise all-atom RMSD values was computed for all structures within a trajectory at 305K (a minimum of 10,300 structures). The number of neighbors of each structure, within a cutoff value of 3.0Å, was counted and the structure with the largest number of neighbors was removed from the pool along with all neighbors. This was repeated on the remaining members of the pool until every structure was assigned to a cluster. Due to memory constraints, each independent replicate was clustered independently and the results were subsequently combined. An examination of pairwise RMSD values between central structures of clusters obtained from different trajectories for the same peptide suggested that a modest reduction in the number of clusters would occur were the data for a given peptide combined before the analysis. The largest effect was seen for central structures obtained from Nt17-Q_5 monomer data with a ~27% reduction.

In the algorithm described above, the number of clusters is not explicitly defined prior to the analysis. The fraction of the ensemble ($F_{ensemble}$) represented by the most populated cluster ranged from 0.215–0.360 for Nt17-Q_N dimers and 0.160–0.312 for monomers and the number of clusters required to represent 95% of the ensemble was equal to 0.78 to 4.32% of the identified clusters for

Nt17-Q_N dimers and 14.8 to 42.2% for monomers. These statistics (see Table 6.c) indicate that the settings chosen for the algorithm are reasonable.

Solvent Accessible Volume Analysis

The solvent-accessible volume of each atom in each glutamine residue was computed according to Equation 3-3. Data were obtained either for simulations using the ABSINTH model at 305K or for simulations in an excluded volume (EV) reference state in which only steric repulsions are present. The latter is a reference model used in previous work to model conformations with large solvent-accessible volumes.[35,36] The difference between the two values (Δ_{SAV}=SAV_{H20}-SAV_{EV}) provides an estimate of the similarity between the two ensembles from the viewpoint of the solvent accessibility of glutamine residues. Data were accumulated every 500 simulation steps and obtained for three chain lengths for constructs of the types Nt17-Q_N and Q_N.

VI.3.3. CD Spectroscopy

Chemically synthesized, uncapped Nt17-peptide in purified form was obtained as a gift from Trevor P. Creamer. We collected CD spectra using a Jasco J715 spectropolarimeter with a 1mm path length quartz cuvette. Nt17 (1mg) was initially dissolved in a mixture of 1ml of trifluoroacetic acid (TFA) and 1ml of hexafluoroisopropanol (HFIP). This solution was divided in half and evaporated under a nitrogen stream followed by lyophilization for one hour to remove any residual TFA and HFIP. The peptide was re-suspended to a concentration of 1mg/mL in freshly prepared solutions of either 6M urea or 50%

(v/v) trifluoroethanol (TFE) in 100mM phosphate buffer (pH 7.2). These stocks were diluted to a final working concentration of 0.1mg/mL (50μM). Wavelength scans were performed at a rate of 50nm per minute in discrete steps of 0.1nm. The temperature was held constant at 25°C. The reported results represent averages of ten scans.

VI.4. Results

VI.4.1. Secondary Structure Propensity of Nt17 as a Function of PolyQ-Expansion Length

Figure 6.1 illustrates that the Nt17 peptide in isolation may well exist as an amphipathic α-helix. It is suggested experimentally and computationally that transient helical segments may form and that their stabilities are dependent on solution conditions.[26,29] However, the secondary structure propensity of polyQ-expanded Nt17 is uncharacterized. We therefore investigate the dependence of the expected α-helical propensity on polyQ-expansion length. Polyglutamine in its soluble form is well-known to be disordered and consequently we also ask whether the Nt17-fragment induces any canonical secondary structure in polyQ-expansions.

Panel A of Figure 6.2 shows that Nt17-Q_N monomers show high α-helicity within the Nt17-stretch for short polyQ-expansions and low temperatures. For temperatures T<365K, this propensity is depleted in a polyglutamine length-dependent manner and approaches the baseline defined by the high temperature limit for N>20. From Panel A we also see that the polyQ-expansion suppresses

the inherent secondary structure propensity of the Nt17 segment. Panel B makes the point that we do in fact observe an induction of α-helix in the Q_N-fragment for small enough N. Similar to the loss of structure in the Nt17-fragment, this trend is N-dependent and vanishes for N>20. This indicates that the entire peptide becomes increasingly disordered with increasing N – a result which makes intuitive sense given the intrinsic preference for disorder exhibited by polyglutamine.

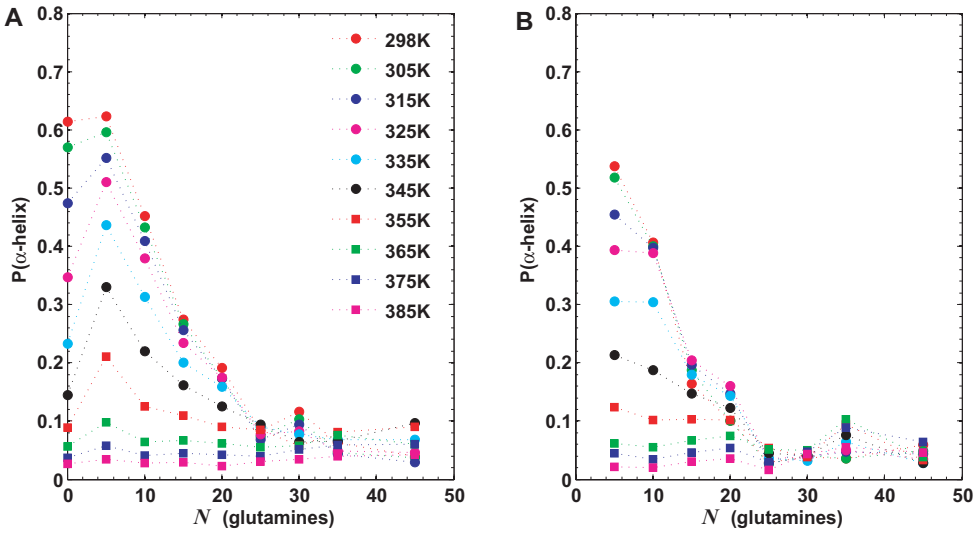

Figure 6.2: α-Helix propensities for constructs of the type Nt17-Q_N. Panel A shows the mean probability of the residues in the Nt17-fragment to adopt α-helical conformation and Panel B shows the same for the Q_N-fragment. Probabilities were calculated by analyzing the backbone ϕ/ψ-angles and considering consecutive residues in the α-helix basin. Only segments of at least three residues in length were counted.

The loss of α-helix propensity in itself is no rigorous indicator of disorder. However, Figure 6.3 shows that the propensity for consecutive residues to populate the β-basin of the Ramachandran map – a prerequisite for the formation

of β-hairpins and β-sheets – is vanishingly small. Both panels show no discernible dependence of the β-strand propensity on the length of the polyQ-expansion. This is consistent with the results obtained in Chapter V which show that polyglutamine is extremely unlikely to form a structure with high β-content under typical conditions. Figure 6.3 rejects a speculative suggestion brought forth in the literature[29] which proposed an induction of β-content in the Q_N-fragment with the Nt17-fragment serving as a putative template.

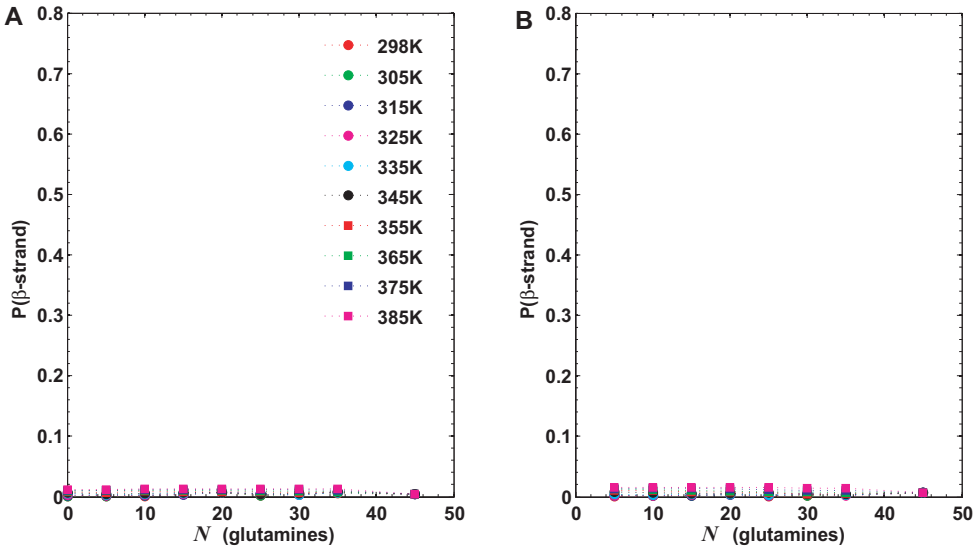

Figure 6.3: β-Strand propensities for constructs of the type Nt17-Q_N. Panel A shows the mean probability of the residues in the Nt17-fragment to form a β-strand and Panel B shows the same for the Q_N-fragment. Probabilities were calculated analogously to Figure 6.2: only here the β-basin of the Ramachandran map is considered.

Based on NMR and CD data, Thakur *et al.*[26] suggest a weak propensity of the Nt17-fragment to form α-helices. However, in the presence of 10% (v/v) TFE,[37,38] significant helicity is induced suggesting that the peptide may be on the

cusp of adopting a stable fold in solution. To corroborate both this result and our own analysis, which indicates high but transient helix-content, we used circular dichroism (CD) spectra as a function of TFE and urea concentration. We expect that with increasing TFE concentration the peptide eventually exists exclusively as a well-defined α-helix. Conversely, we expect that with increasing urea concentration the peptide eventually exists exclusively as a disordered coil. These two limits define the baselines of a two-state model which is analyzed as follows. Even though the exact origin of the CD signal is poorly understood, we assume that the total helix-content is directly proportional to the ellipticity at 222nm. We can then define the fractional α-helix content as:

$$f_\alpha = \frac{\theta_{222}^{H_2O} - \theta_{222}^{coil}}{\theta_{222}^{helix} - \theta_{222}^{coil}} \qquad (6\text{-}1)$$

In Equation 6-1, θ_{222}^{coil} is the baseline value obtained for high urea concentration, and θ_{222}^{helix} is the baseline value obtained at high TFE concentrations. The experimental values were $2.0 \cdot 10^3 \mathrm{deg\,cm^2 dmol^{-1}}$ and $-3.3 \cdot 10^4 \mathrm{deg\,cm^2 dmol^{-1}}$, respectively. Figure 6.4 shows our results for $\theta_{222}^{H_2O}$. The observed dependence on TFE concentration confirms the result of Thakur *et al*. cited above. However, the urea data indicate that there is signficant residual α-helix content in neat buffer conditions, which we estimate to be about 34%. As a comparison and a direct test of the robustness of our analysis following Equation 6-1, we estimated f_α using an empirical reference state proposed in the literature:[39,40]

$$f_{\alpha}^{emp} = \frac{\theta_{222}^{H_2O}}{\theta_{222}^{helix}}$$
$$\theta_{222}^{helix} = -4\times10^{4}\cdot\left(1-\frac{2.5}{N_{pep}}\right)\ \left[\deg\,\mathrm{cm}^{2}\mathrm{dmol}^{-1}\right] \qquad (6\text{-}2)$$

In Equation 6-2, N_{pep} is the length of the polypeptide under investigation. The resultant value for θ_{222}^{helix} is -3.41·10^{4}deg cm^{2}dmol^{-1}, very close the value we obtained from the TFE titration (3.5% difference). Using Equation 6-2, we determine f_{α}^{emp} to be ~29%, also in good agreement with the estimate from Equation 6-1.

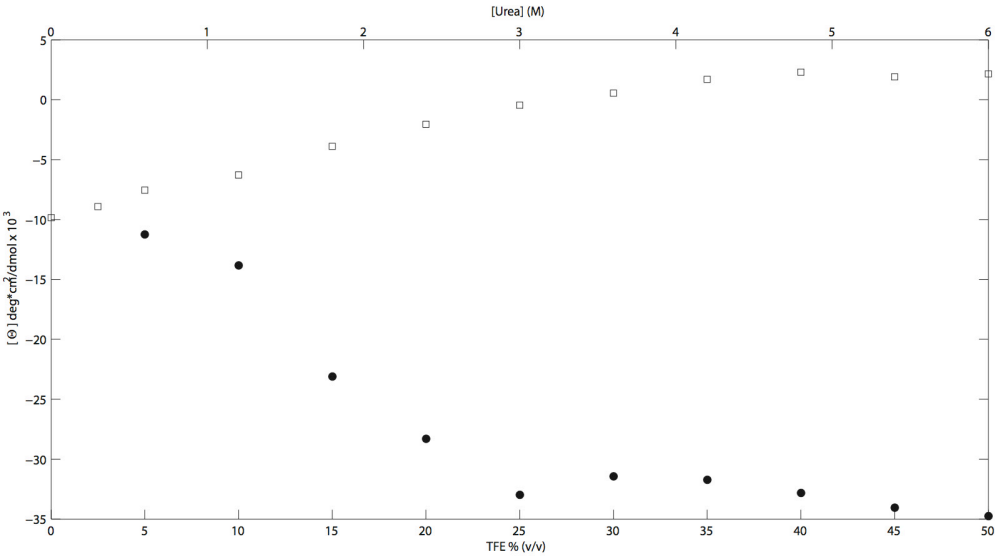

Figure 6.4: CD spectroscopic analysis of the α-helix propensity of the Nt17-peptide. Data were obtained at 298.15K and as a function of TFE (filled circles) and urea (open circles). Helicity is calculated from the data as detailed in the text.

Comparison to Figure 6.2 reveals that the computational estimates for the α-helix content suggested higher values: for the isolated Nt17 peptide at 298K we obtain a value of 61.4%. There are two possible explanations: i) the

computational ensemble exhibits artificially enhanced helix content, and ii) the computational and experimental readouts are incongruent. While we cannot exclude i), results in Chapter III show that the helix-coil transition is quantitatively well-described by the ABSINTH model.[6] We address ii) in Figure 6.5. Panel A shows that the estimate of helicity for the isolated Nt17 fragment of 33.5% agrees with the CD estimate if we stipulate that at least two helix turns (seven residues) need to be formed in order to observe helicity in CD experiments. Qualitatively, the trends established in Figure 6.2 are all preserved with this modified definition.

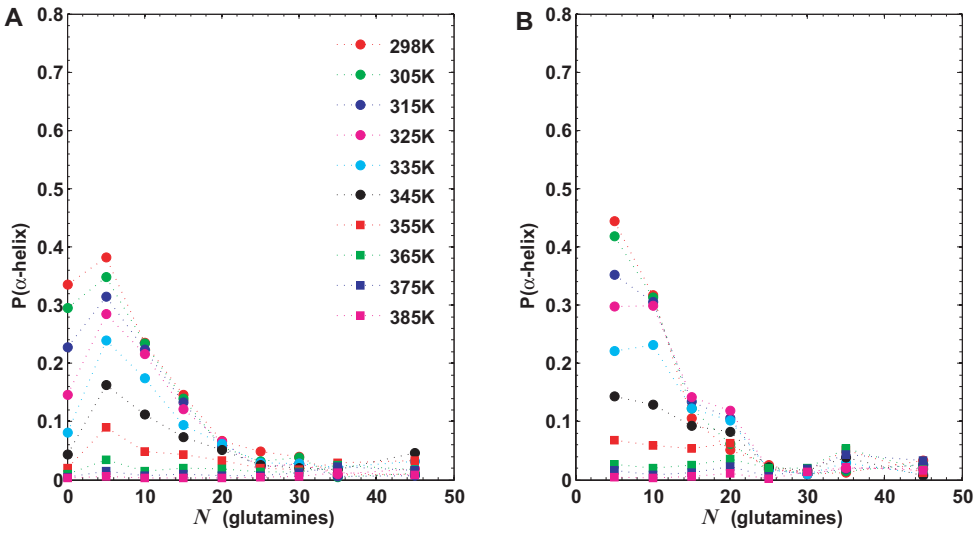

Figure 6.5: Adjusted α-Helix propensities for constructs of the type Nt17-Q_N. These plots are identical to those in Figure 6.2 with the exception that only segments of at least seven consecutive residues in α-helical conformation were considered.

VI.4.2. Polymeric Properties of Chimeric Peptides

In previous work (see Chapters II[36] and IV[4]) we established that polyglutamine intrinsically prefers to collapse to disordered globules in poor solvent conditions and that it undergoes a well-defined globule-to-coil transition

with increasing solvent quality which can be modulated by simulation temperature. Panels B of Figures 6.2 and 6.3 suggest that this intrinsic preference might be preserved. Figure 6.6 shows that as a function of temperature the chimeric Nt17-Q_N-peptides swells much like the homopolymer does:

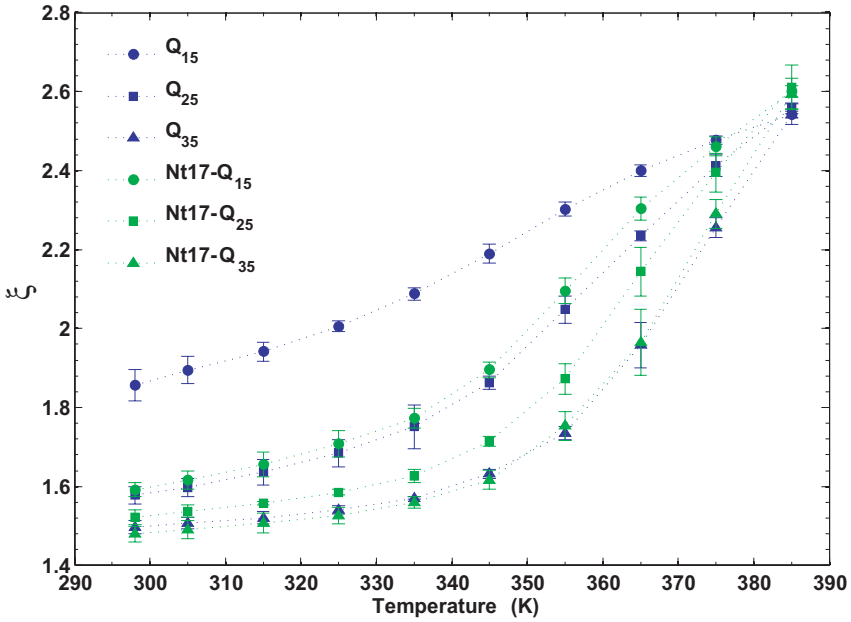

Figure 6.6: Coil-to-globule transition for chimeric peptides. The average radius of gyration was calculated and normalized by chain length. The θ-temperature, T_θ, was estimated to be ~390K in previous work.[5] Consistent with that estimate, we find that T_θ may be set to 385K for this work based on the intersection of all curves at that temperature. ξ is defined as $\langle R_g \rangle / \sqrt{Z_{pep}}$ where Z_{pep} is the total peptide length.

A comparison of data for Nt17-Q_{25}, Nt17-Q_{15}, and Q_{25} in Figure 6.6 shows that from a coarse polymer-centric view the latter two peptides behave nearly indistinguishably while Nt17-Q_{25} shows a significantly different swelling profile. In such a scenario, the global response to changes in solvent quality does not seem

to depend strongly on the precise nature of the polypeptide sequence, but more on overall peptide length. Conversely, data for Nt17-Q_{35} and Q_{35} appear perfectly matched indicating that the effects of overall length and polyQ-expansion length are not independent of one another. This is a very important consideration for the design of experimental controls: it questions for example whether the modulating effects of the Nt17-fragment on the peptide Nt17-Q_{15} can be adequately tested by comparing its properties to those of Q_{15}.

If from a polymeric standpoint the chimeric peptides behave similarly to the homopolymers, we may conjecture that the Nt17-fragment becomes more extended in the presence of long enough polyQ-expansions. This would be driven by the entropic gain in forming diverse interfaces with other parts of the polymer showing little overall specificity. This idea is supported by the analysis of contact distributions shown further below and suggested by the loss of secondary structure propensity established in VI.4.1. An N-dependent swelling of the Nt17-fragment in chimeric peptides is shown by Thakur *et al*. via the use of a suitable FRET pair.[26]

Figure 6.7 shows that we do see such a transition in the simulations as well. The ensemble-averaged radius of gyration (R_g) computed for just the Nt17-fragment in Nt17-Q_N monomers shows the adoption of a collapsed conformation for small N at physiological temperatures. This is clear from the fact that the values are below those obtained for a straight α-helix which in itself is a rather compact conformation for a peptide of this length. This indicates that the high helix-content observed in Figure 6.2 is not the result of the peptide forming a

straight α-helix, but rather a result of the population of transient, shorter helix segments consistent with the data shown in Figure 6.4. The increase in R_g as a function of N indicates swelling induced by the presence of the polyQ-expansion. This signal correlates with the loss of α-helix content seen in Figures 6.2 and 6.4. The observation of the swelling of Nt17 as a function of N has important consequences for the formation of an interface between the two segments and this is discussed next.

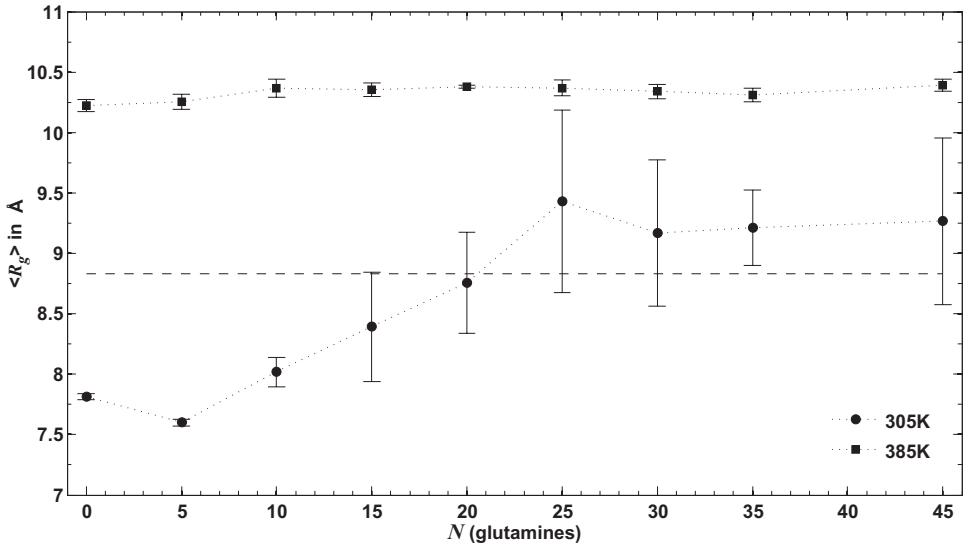

Figure 6.7: Radius of gyration of the Nt17-fragments in chimeric peptides at 305K and 385K. The dashed line represents the R_g of just the Nt17-fragment in a straight α-helical conformation (8.83Å, compare Figure 6.1).

VI.4.3. Characterization of Intra- and Intermolecular Interfaces Formed by Chimeric Peptides

In VI.4.1 and VI.4.2 we demonstrated that, for large enough N, peptides of the type Nt17-Q_N have no discernible secondary structure preferences and that

the Nt17-fragment becomes more extended. We stipulate that this is consistent with the Nt17-segment forming an extended interface with the Q_N-segment, and test this idea by quantitative analysis of contact probabilities. Panel A of Figure 6.8 shows intramolecular contact probabilities (see VI.3.2) for monomeric chimeras at 305K:

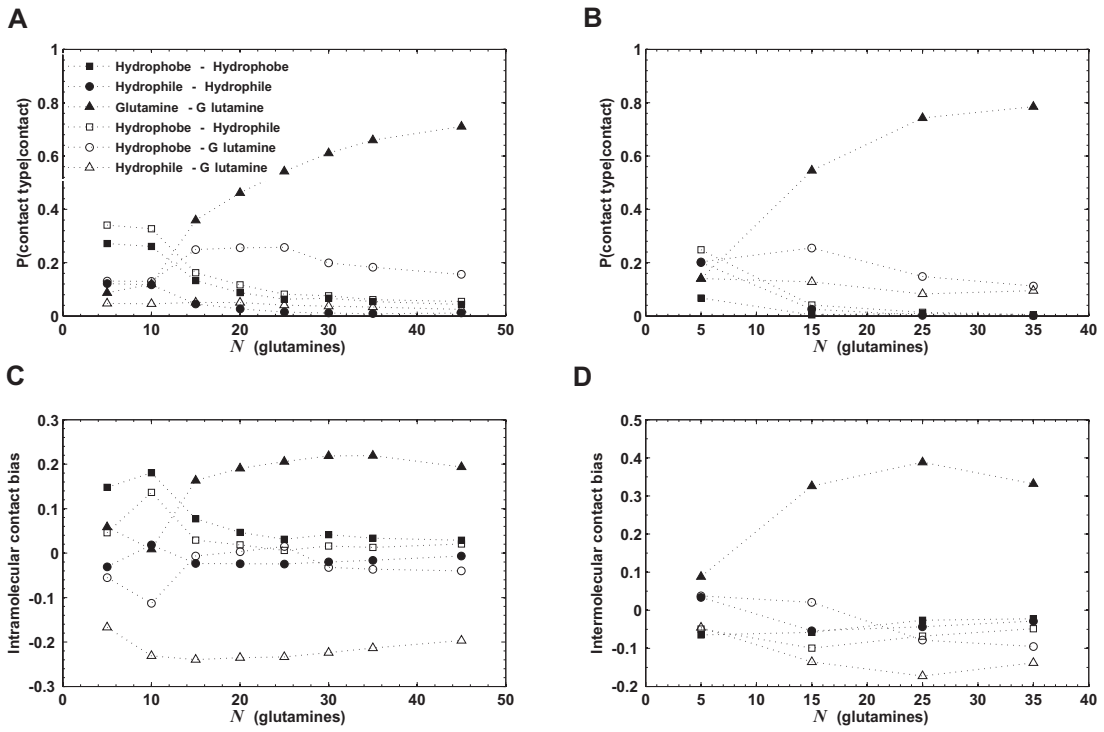

Figure 6.8: Intra- and intermolecular contact probabilities for chimeric peptides. The length dependence of the probability of observing a contact between specific classes of residues in Nt17-Q_N constructs is shown. The definition of a contact and the calculation of a random prior model are provided in VI.3.2. The contact bias is computed as the difference of the observed contact probability from that of a random prior. Panel A shows intramolecular contacts in Nt17-Q_N monomers at 305K, Panel B intermolecular contacts in Nt17-Q_N dimers at 305K, Panel C intramolecular and Panel D intermolecular contact biases.

The data in Panel A indicate three defining features of the intramolecular interactions within Nt17-Q_N: (i) intra-segment Nt17 contacts are significantly present only at short polyQ-expansion lengths (N<15), (ii) the polyglutamine segment retains the ability to form extensive collapsed interactions as glutamine-glutamine contacts prevail when N increases, and (iii) an intramolecular interface forms between the Nt17-hydrophobes and the polyglutamine segment. Panel C reveals that the data are partially explainable by combinatorial arguments: by subtracting from the observed probabilities those of a random prior (see VI.3.2), we establish that the "excess" interactions for N>15 are dominated by a promotion of glutamine-glutamine contacts and a depletion of hydrophile-glutamine interactions. For smaller N, hydrophobes play a much more pronounced role: Panel C reveals that the dominance of hydrophobe-hydrophile interactions seen in Panel A for small N is mostly explained by combinatorial arguments and that it really is hydrophobes which associate with each other in a preferential manner. It should be noted that the random prior does not correct for proximity relations in the primary sequence with the exception of the exclusion of nearest-neighbor contacts.

For a select set of chimeric peptides, we simulated their intermolecular association as detailed in VI.3.1. As is shown below (see Figure 6.11), the molecules associate at low enough temperatures and intermolecular contacts are observed often enough for analyses to be quantitative. In Panels B and D of Figure 6.8 we apply the same analysis as in Panels A and C to intermolecular contacts within Nt17-Q_N dimers. Our data demonstrate that the intermolecular

interface is dominated by glutamine-glutamine contacts. Interactions of hydrophobic residues with hydrophobes or glutamine residues on the other chain play a dominant role only for the shortest peptide studied. The comparison to the combinatorial prior in Panel D demonstrates that the interface is predominantly formed across the polyglutamine segments of two monomers. This is a rather striking signature since all other contact types appear indifferent or diminished. These data are consistent with the observation of the formation of an intramolecular interface between the Nt17-hydrophobes and the polyglutamine segment in Nt17-Q_N monomers. We speculate that only those hydrophobes that are not sequestered into an intramolecular interface may contribute to the formation of the intermolecular interface.

Our previous results presented in Chapter IV[4] suggest that polyglutamine is strongly associative, and this is consistent with the observation that contacts between residues of the Nt17-segments of each monomer do not contribute significantly to the formation of the dimer interface. Conversely, it has been proposed in the literature that the intermolecular interface forms across the hydrophobic face of the Nt17-segment[29] and that the hydrophobic residues in the Nt17-segments are crucial determinants in mediating associativity.[26] The latter conclusion comes from the profound impact single-point mutations showed on the aggregation propensity of chimeric peptides. Our data challenge the mechanistic interpretation that this points to an association mechanism in which the Nt17-segments provide the intermolecular interface.

So far the data suggest that the structural impact of the Nt17-fragment is relatively minor for N>15. We next asked whether there are dominant members in the ensembles characterized by a specific topology or structure which eluded the analysis thus far. Moreover, we wished to quantify whether the ensembles of dimer structures are composed of structures observed in the monomer ensembles and – by extension – whether the presence and nature of higher-order oligomers may be predicted from the data presented here. Table 6.c summarizes the statistics of a clustering of all conformations (see VI.3.2):

	Monomer			Dimer		
System	**$N_{clusters}$**	**$F_{ensemble}$**	**N_{95}**	**$N_{clusters}$**	**$F_{ensemble}$**	**N_{95}**
Nt17-Q_5	567	0.160	239	3326	0.250	81
Nt17-Q_{15}	169	0.297	53	797	0.215	29
Nt17-Q_{25}	46	0.312	18	1537	0.360	12
Nt17-Q_{35}	61	0.278	21	278	0.297	12

Table 6.c: Cluster statistics for ensembles of Nt17-Q_N. $N_{clusters}$ denotes the total number of clusters found for the specified ensemble and $F_{ensemble}$ indicates the fraction of structures contained within the most populated cluster for that ensemble. N_{95} is the minimum number of clusters needed to represent 95% of the entire ensemble of structures given the fixed cluster assignment. For dimers, $F_{ensemble}$ and N_{95} are restricted to those clusters representing associated states as defined by a minimum distance criterion (<3.0Å).

From Table 6.c, it does appear as if a large fraction of the data may be represented by a small number of clusters. This point has to be considered

carefully, however, since poor sampling would give rise to the same trend and this may be the case for the longer peptides. Nonetheless, it does appear relatively clear that no specific fold is adopted for any of the peptide constructs studied as indicated by the numbers for N_{95} which are significantly larger than unity. The central structure taken from the most populated cluster from each monomer ensemble is presented in Figure 6.9:

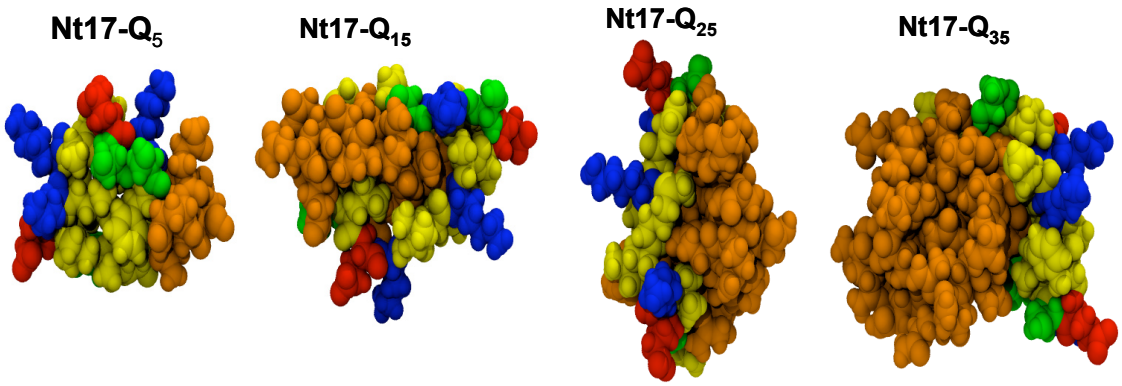

Figure 6.9: Central structures of the most populated cluster for monomeric chimeras. Graphics were generated using VMD.[30] Atoms are drawn in space-filling representation and colored according to residue type: positively charged (blue), negatively charged (red), hydrophilic excluding glutamine (green), hydrophobic (yellow), and glutamine (orange).

Figure 6.9 suggests that all monomer structures share a common feature where the charged groups are separated from both the glutamine and the hydrophobic residues. For Nt17-Q_5, the Nt17-fragment is collapsed unto itself and the hydrophobic residues form a semi-accessible core which forms an interface with the glutamine residues. This is consistent with the analysis in Panels A and C of Figure 6.8. For larger N, the Nt17-fragment becomes more

extended (Figure 6.7) and the hydrophobic residues are sequestered against the polyglutamine domain. Clearly, the primary sequence determines that the hydrophobic residues remain in the vicinity of the charged residues and – by extension – close to the surface of the globular structure. This topological frustration suggests that they play little role in the formation of soluble oligomers, for which the "micellar" nature, *i.e.*, the exposure of the charged groups to the solvent, is preserved. Their role might be much more amplified, however, under conditions in which desolvation of the charged amino acid sidechains is possible, *i.e.*, at the onset of phase separation. Such a scenario could reconcile the apparent conflict with experimental data stated in the discussion to Figure 6.8.

The average topology observed in Figure 6.9 allows the prediction that soluble oligomers will form in a way that preserves the monomers' "micellar" character. Figure 6.10 shows representations of the central structure of the dominant clusters for associated dimers. It confirms what Figure 6.8 reported, *i.e.*, that the association is mediated by dominant glutamine-glutamine contacts augmented by glutamine-hydrophobe contacts for small N. For the two longest peptides shown, the Nt17-fragment is significantly extended and forms a band across the large, globular, dimeric polyglutamine domain. All structures in Figure 6.10 exhibit interfaces which appear amenable to the formation of higher-order oligomers. However, they do suggest that the distribution and structure of the populated oligomers in solution may be very different from what is seen for the homopolymer due to the constraint of maintaining a global, "micellar" architecture. The elucidation of important oligomer sizes and topologies is the

subject of ongoing and future work. Interestingly, the Alzheimer's peptide Aβ42, just like Nt17-Q_N a heteropolymeric and amyloidogenic system, is known to form soluble oligomers of specific sizes.[41,42]

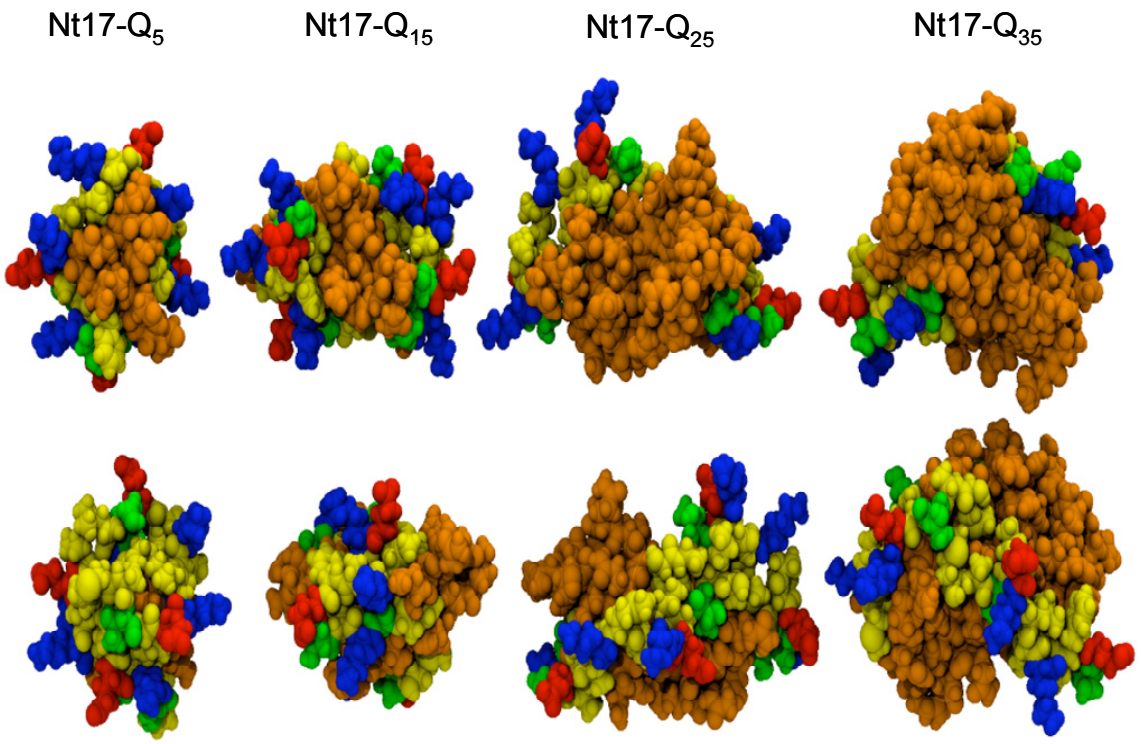

Figure 6.10: Central structures of the most populated cluster for associated chimeras. Graphics were generated using VMD.[30] The bottom row is the same structure as the top row rotated by 180° around a horizontal in-plane axis. Atoms are drawn identically to Figure 6.9.

VI.4.4. Associativity of Chimeric Peptides in Comparison to Different Control Peptides

Previously, we found that dimerization of polyglutamine homopolymers occurs spontaneously at room temperature and effective concentrations approximating those of a typical *in vitro* experiment.[4,5] Here, we probe the effect

of the Nt17-segment on spontaneous dimer formation in polyQ-expanded peptides. To quantify associativity we computed a temperature-dependent excess interaction coefficient $B_{22}(T)$ identical to previous work (see Equation 4-3), which can be viewed as a normalized second virial coefficient. The θ-temperature was 385K as shown in Figure 6.6. The reader is reminded that $B_{22}(T)$ is a saturating parameter with a well defined lower (always associated) bound. Panel A of Figure 6.11 plots $B_{22}(T)$ as a function of polyQ-expansion length for dimers of:

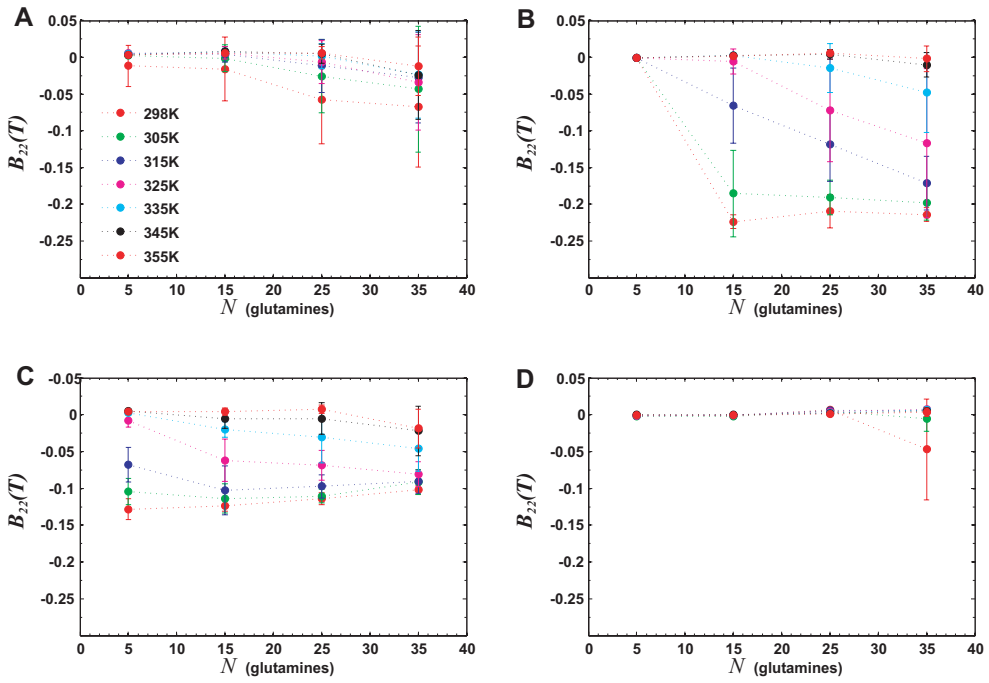

Figure 6.11: The associativity of chimeric peptides relative to controls. The temperature-dependent excess interaction coefficient $B_{22}(T)$ as a function of polyglutamine chain length from simulations of two chains is plotted for Nt17-Q_N at an effective concentration of 100μM in Panel A, for Q_N at an effective concentration of

100μM in Panel B, for Nt17-Q_N at an effective concentration of 500μM in Panel C, and for $K_2Q_NK_2$ at an effective concentration of 100μM in Panel D.

The data in Panels A and B indicate that the associativity of polyglutamine-expanded chains is significantly reduced by the presence of the Nt17-fragment when compared to the homopolymers of identical glutamine-length. Even though the net associativity is weak, similar conclusions with respect to the N- and T-dependencies of $B_{22}(T)$ as before are obtained (see Chapter V).[5] Panel C clarifies that the large attractive interfaces postulated based on Figure 6.9 are not misleading: if the concentration is increased fivefold, strong associativity is observed for all chimeric constructs. The absolute values of $B_{22}(T)$ are more positive simply because the droplet size is smaller (see Equation 4-3). Analysis of cumulative distribution functions revealed that the chains are mostly associated even for polyQ-expansions as short as five residues (data not shown).

Thakur *et al*.[26] show a profound enhancement of the bulk aggregation rate of both Nt17-$Q_{35}K_2$ and Nt17-$K_2Q_{36}K_2$ over $K_2Q_{35}K_2$. Panel D of Figure 6.11 shows that our data are not necessarily inconsistent with this experimental result. When we use $K_2Q_NK_2$ as the control, aggregation of the "homo"polymer is significantly reduced compared to the chimeric peptide shown in Panel A. This result emphasizes again the need for careful controls if one is interested in delineating the *intrinsic* effects polyglutamine on the aggregation of polyQ-expanded peptides.

The somewhat surprising lack of associativity exhibited by $K_2Q_NK_2$ prompted us seek a structural explanation for this phenomenon. Figure 6.12 is analogous to Figure 6.10 and shows the central structure of the most populated cluster for the only peptide forming appreciable dimers at 100μM, K_2-Q_{35}-K_2. Figure 6.12 shows that the structures of dimers that do form must associate in a manner that separates the terminal K_2-groups on each chain and keeps them solvent-exposed. This suggests that the associativity is weaker than that of the true homopolymer or of the chimeras due to the increased difficulty in finding a productive dimer conformation given the K_2-group repulsion.

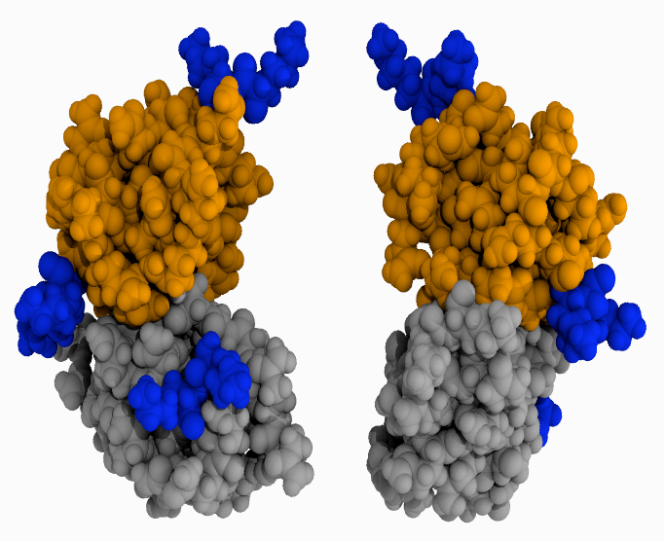

Figure 6.12: Central structures of the most populated cluster for associated $K_2Q_{35}K_2$. Graphics were generated using VMD.[30] The right panel is rotated by 180° around a vertical axis. Atoms are drawn identically to Figure 6.9 with the exception that the glutamine residues of the second molecule are shown in grey.

Can we quantitatively identify an origin for the negative impact the Nt17-segment has on the intrinsic associativity of polyglutamine? In Figure 6.13, we

compare the change in solvent accessible volume from an excluded volume (EV) reference state (Δ_{SAV}):

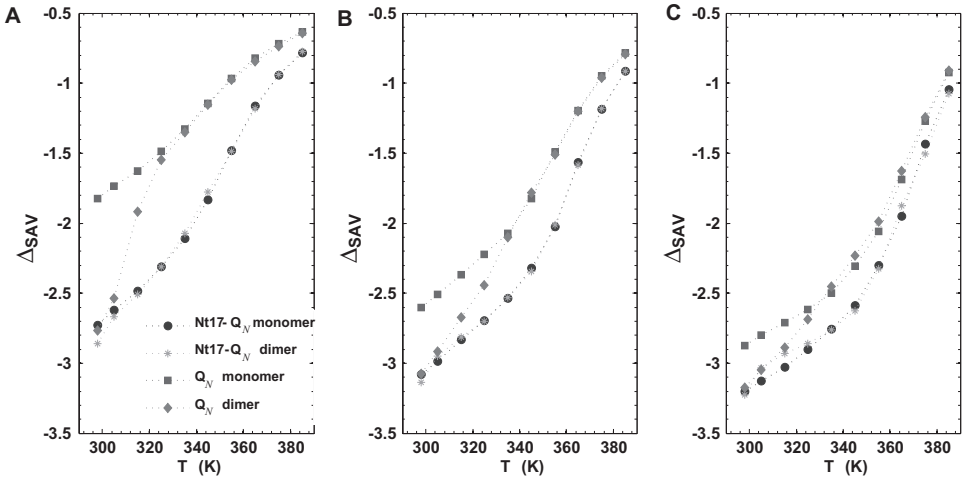

Figure 6.13: Change in solvent accessible volume per glutamine residue for chimeric peptides from the EV reference state. Data are plotted as a function of simulation temperature for the chimeric peptides Nt17-Q_N and their homopolymeric counterparts Q_N for lengths N of 15 in Panel A, 25 in Panel B, and 35 in Panel C. Data are shown for both the monomer and the dimer case. For details on the calculation of Δ_{SAV} see VI.3.2. Note that values generally become more negative with increasing chain length; this indicates that sequestration from solvent becomes more and more complete for larger peptides.

Figure 6.13 shows that the values are in general negative: the simulated ensembles in water are less accessible to solvent than the EV reference state. They become more negative with decreasing temperature indicative of collapse in both the monomer and dimer cases. However, the data make the point that Q_N-homopolymers achieve significant sequestration of glutamine residues from the solvent by forming dimers. This is indicated by the more negative values

seen for dimers of Q_N at temperatures at which association occurs when compared to monomers (see Figure 6.11). This trend is universally observed for all three polyglutamine lengths shown. Conversely, for Nt17-Q_N monomers glutamine residues are just as effectively sequestered from the surrounding solvent as they are for associated Nt17-Q_N dimers. This is an intriguing result and suggests that the association-prone interface, *i.e.*, the polyglutamine segment, is partially sequestered by the Nt17-segment and hence protected from intermolecular association thereby explaining the reduced associativity observed in Figure 6.11.

Based on Figure 6.12, we conjecture that the "reactive" interaction surface of the chimeric peptides is much smaller than that of the homopolymers. In energetic terms, this would correspond to a much reduced "surface tension" of the chimeric peptides. We carry out a decomposition of system energies similar to previous work (see Equation 4-4):

$$\frac{\left\langle U_{total}(T) - U_{Nt17}(T) \right\rangle}{N} = C_1(T) + C_2(T) N^{-1/3} \qquad (6\text{-}3)$$

In Equation 6-3, U_{total} is the total system energy and U_{Nt17} is the total system energy of the isolated Nt17-peptide including the counterions. For application to polyglutamine homopolymers, the term U_{Nt17} is ignored. Equation 6-3 is written in such a way that the internal contribution of the Nt17-fragment would be normalized out completely if its conformation remained rigid for the entirety of the simulation. However, Equation 6-3 does include contributions from

both changes in conformation of the Nt17-segment and from its interactions with the polyQ-expansion.

Figure 6.14 shows the results we obtain for monomeric peptides. The analysis in Figure 6.14 indicates that the Nt17-segment alters the energetics of the Q_N-segment profoundly. In the collapse regime, a significant increase in the volumetric term of ~5kcal/mol relative to the homopolymer indicates a reduction of self-interactions within the polyQ-expansion (Panel A). This confirms that the Q_N-segment does not adopt as tightly a globular conformation as it would in the homopolymer.

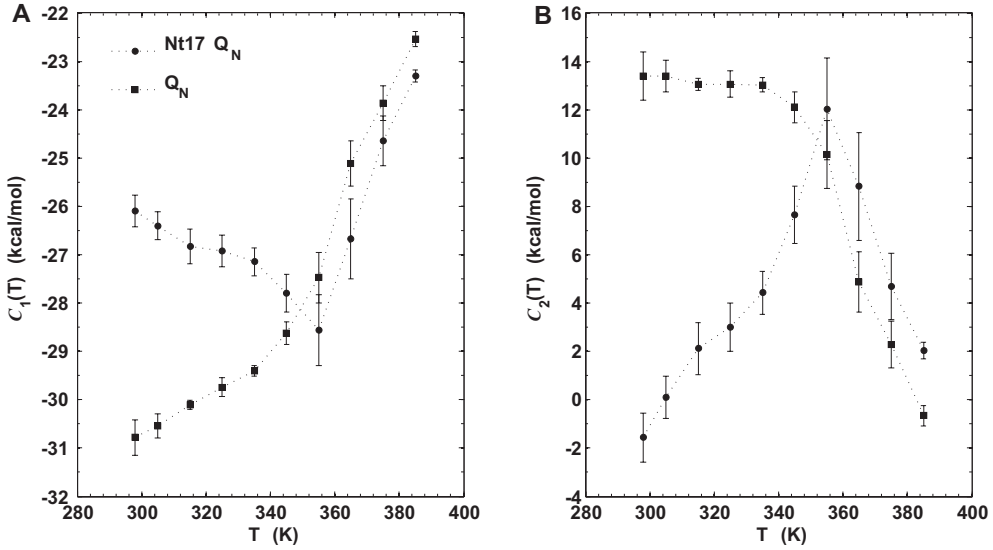

Figure 6.14: Volumetric and surface energy contributions for chimeric and homopolymeric peptides. Volumetric (C_1 shown in Panel A), and surface (C_2 in Panel B) energy contributions obtained from fits according to equation 6-3 for Nt17-Q_N and Q_N monomers. Fits were obtained from data for chain lengths N=(15, 20, 25, 30, 35, 45). The quality of the fits underlying these data is assessed through the plotted errors bars which result from a rigorous propagation of error though the fitting procedure (see

VI.3.2). Only chain lengths of N≥15 are used, as shorter chains are not likely to have a volumetric contribution to the potential energy.

Do the more positive values for C_1 imply that globules formed by the chimeric peptides are less stable overall? The analysis in Figure 6.6 contradicts such a claim since T_θ obtained for homopolymeric polyglutamine appears to be unaltered by the presence of the Nt17-flanking sequence. Moreover, Panel B shows that the surface energy penalty experienced by the polyglutamine domain is significantly reduced in the presence of the Nt17-fragment. This is indicative of the formation of a favorable interface between the two pieces of the chimeric peptide. Conversely, the homopolymer exhibits a large surface penalty indicating that globules of Q_N are highly amenable to intermolecular association. Panel B of Figure 6.14 therefore gives a direct explanation for the reduced associativity of the chimeric peptides seen in Figure 6.11.

With increasing temperature, the discrepant behavior of the two different systems becomes less pronounced, and the data roughly agree from 350K onward. We interpret this to mean that with increasing temperature the polyQ-segment becomes gradually dissociated from the Nt17-segment as the driving for overall collapse is weakened. In the absence of an extensive interface, the contributions from the Nt17-fragment are normalized out by Equation 6-13, and the curves for the two different peptide constructs should and do agree.

VI.5. Summary and Discussion

In this chapter, we have investigated the impact of the 17 residues N-terminal to the polyQ-expansion in huntingtin on the intrinsic properties of polyglutamine. The relevance of this study lies in the issue of sequence context dependencies of exonic CAG repeat diseases: what are representative *in vivo* fragments, and how may the presence of flanking sequences explain differences observed in pathogenesis? HD is the most prevalent of the exonic CAG repeat diseases and exon1 of the host protein possesses a fragment N-terminal to the polyQ-expansion which was shown to increase aggregation rates relative to specific control peptides.[26] Furthermore, peptides carrying the Nt17-segment showed different subcellular localization and a different propensity to form visible aggregates in cellular models of disease (see VI.2). Our results may be summarized as follows:

- The intrinsic preference of the Nt17-fragment to form an α-helix becomes increasingly suppressed in the presence of polyQ-expansions of increasing length. Specifically, for N>15 the chimeric peptides as a whole form disordered globules at physiological conditions (Figures 6.2 to 6.6).
- The Nt17-fragment undergoes a distinct swelling transition and forms an interface of increasing size with increasing N (Figures 6.7 and 6.9). Intramolecular contacts are dominated by glutamine-glutamine interactions for large N but a significant interface between the hydrophobic residues in the

Nt17-segment and the polyQ-domain is observed (Panels A and C of Figure 6.8).

- We find that chimeric peptides form dimers *less* readily than homopolymers but *more* readily than experimental control peptides carrying flanking lysine residues for identical N (Figure 6.11). We show that the presence of the Nt17-fragment imposes a “micellar” nature on the accessible conformation space (Figure 6.10) which restricts productive encounters.
- We establish quantitatively that the dimerization propensity is mediated by the polyglutamine stretch (Panels B and D of Figure 6.8). It is quenched in the chimera because the polyQ-expansion forms an extended and favorable interface with the Nt17-fragment (Figures 6.10 and 6.13) that lowers the effective “surface tension” of the polymer (Figure 6.14).

Our results have a significant impact on the understanding of the effects of flanking sequences on the molecular properties of polyglutamine. The cross-talk of secondary structure propensities we observe points to the protective effect certain flanking sequences may exhibit. If a long enough polyQ-expansion is found in the context of a well-folded protein, it may disrupt the fold and cause additional stress to the PQCS. If, however, the expansion is short enough that the host protein domain induces regular secondary structure in the polyQ-segment, sequence context may relieve its deleterious effects. Finally, if the polyQ-stretch is not part of a well-folded domain, it can be expected to adopt conformations corresponding to the intrinsic preferences of polyglutamine and sequence context might be less important. While evidence for differential effects

in line with the first two scenarios is found in the literature,[43,44] the last scenario is very much consistent with the popular toxic fragment hypothesis (see I.2.4 and I.2.7). Here, the fragments would eventually be short enough that the properties of polyglutamine override the preferences of the remaining flanking residues. From such a model, a reasonable explanation for the astute, but variable length dependence of the disease age-of-onset[45] (see Table 1.a) begins to emerge.

The modulation of associativity that we observe is contradicting the work by Thakur *et al.*[26] at first glance only. With respect to control peptides solubilized by the addition of lysine residues, we observe the same trend that is seen experimentally. We also report a similar swelling of the Nt17-fragment with increasing polyglutamine length. Interestingly, electron micrographs show that early stages of aggregation of chimeric peptides are characterized by the presence of spherical oligomers.[26] These oligomers precede the formation of fibrillar aggregates which remain characterized by substantial polymorphism. Our results suggest that the reactive surface for oligomer formation is glutamine-rich. We therefore speculate that higher-order oligomers form along glutamine-glutamine and mixed hydrophobe-glutamine interfaces but that the "micellar" nature imposed by the presence of charged groups will restrict oligomer formation in an astutely N-dependent manner. Alternatively – as suggested by Lee *et al.*[46] – large, linear aggregates may form in solution. Our interpretation of these data from the point of view of this chapter would be that the flanking lysine residues impose this topology at the level of not only the dimers (Figure 6.12) but also larger aggregates. We disagree with the proposed mechanism by which the

amphipathic nature of the Nt17-segment is a mediator of intermolecular associations for oligomer formation.[26,29] Our results suggest that topological frustration imposed by primary sequence lessens the effect of hydrophobic residues in comparison to the polyQ-expansion.

In summary, our results largely agree with what is known about this system *in vitro*. Specific localization effects and an association with mitochondria were observed *in vivo* for chimeric peptides carrying the Nt17-fragment. At this point, our data offer the insight that the Nt17-fragment – due to the micellar architecture of monomers and dimers – remains exposed to the surface of whatever glutamine-rich fragment it is attached to. It appears plausible that the surface-exposed location may predispose the Nt17-fragment to engage in transient interactions much like a functional IDP. Protein-protein or protein-membrane interactions are a minimal requirement for a species to act as a localization signal, and our data explain why those interactions may not be obliterated despite the formation of soluble oligomers.[47]

VI.6. Bibliography

1. Bhattacharyya, A.; Thakur, A. K.; Chellgren, V. M.; Thiagarajan, G.; Williams, A. D.; Chellgren, B. W.; Creamer, T. P.; Wetzel, R. J Mol Biol 2006, 355(3), 524-535.

2. Darnell, G.; Orgel, J. P. R. O.; Pahl, R.; Meredith, S. C. J Mol Biol 2007, 374(3), 688-704.

3. Thakur, A. K.; Wetzel, R. Proc Natl Acad Sci U S A 2002, 99(26), 17014-17019.

4. Vitalis, A.; Wang, X.; Pappu, R. V. J Mol Biol 2008, 384(1), 279-297.

5. Vitalis, A.; Lyle, N.; Pappu, R. V. Biophys J 2009, *in press*.

6. Vitalis, A.; Pappu, R. V. J Comput Chem 2009, 30(5), 673-699.

7. Jacob, J.; Baker, B.; Bryant, R. G.; Cafiso, D. S. Biophys J 1999, 77(2), 1086-1092.

8. Eberhardt, E. S.; Loh, S. N.; Hinck, A. P.; Raines, R. T. J Am Chem Soc 1992, 114(13), 5437-5439.

9. Reimer, U.; Scherer, G.; Drewello, M.; Kruber, S.; Schutkowski, M.; Fischer, G. J Mol Biol 1998, 279(2), 449-460.

10. Hinderaker, M. P.; Raines, R. T. Protein Sci 2003, 12(6), 1188-1194.

11. Williamson, T. E.; Vitalis, A.; Crick, S. L.; Pappu, R. V. Nat Struct Mol Biol 2009, *to be submitted*.

12. Zhang, H.; Li, Q.; Graham, R. K.; Slow, E.; Hayden, M. R.; Bezprozvannya, I. Neurobiol Dis 2008, 31(1), 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.; Ellerby, L. M. Neurobiol Dis 2006, 21(2), 381-391.

14. Nozaki, K.; Onodera, O.; Takano, H.; Tsuji, S. Neuroreport 2001, 12(15), 3357-3364.

15. 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.; Richards, W. G. Neurobiol Dis 2002, 11(1), 111-122.

16. 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.; Ellerby, L. M. J Biol Chem 2007, 282(41), 30150-30160.

17. Gafni, J.; Hermel, E.; Young, J. E.; Wellington, C. L.; Hayden, M. R.; Ellerby, L. M. J Biol Chem 2004, 279(19), 20211-20220.

18. Ellerby, L. M.; Andrusiak, R. L.; Wellington, C. L.; Hackam, A. S.; Propp, S. S.; Wood, J. D.; Sharp, A. H.; Margolis, R. L.; Ross, C. A.; Salvesen, G. S.; Hayden, M. R.; Bredesen, D. E. J Biol Chem 1999, 274(13), 8730-8736.

19. Haacke, A.; Broadley, S. A.; Boteva, R.; Tzvetkov, N.; Hartl, F. U.; Breuer, P. Hum Mol Genet 2006, 15(4), 555-568.

20. Kaltenbach, L. S.; Romero, E.; Becklin, R. R.; Chettier, R.; Bell, R.; Phansalkar, A.; Strand, A.; Torcassi, C.; Savage, J.; Hurlburt, A.; Cha, G. H.; Ukani, L.; Chepanoske, C. L.; Zhen, Y. J.; Sahasrabudhe, S.; Olson, J.; Kurschner, C.; Ellerby, L. M.; Peltier, J. M.; Botas, J.; Hughes, R. E. Plos Genetics 2007, 3(5), 689-708.

21. Rockabrand, E.; Slepko, N.; Pantalone, A.; Nukala, V. N.; Kazantsev, A.; Marsh, J. L.; Sullivan, P. G.; Steffan, J. S.; Sensi, S. L.; Thompson, L. M. Hum Mol Genet 2007, 16(1), 61-77.

22. Orr, A. L.; Li, S.; Wang, C. E.; Li, H.; Wang, J.; Rong, J.; Xu, X.; Mastroberardino, P. G.; Greenamyre, J. T.; Li, X. J. J Neurosci 2008, 28(11), 2783-2792.

23. Steffan, J. S.; Agrawal, N.; Pallos, J.; Rockabrand, E.; Trotman, L. C.; Slepko, N.; Illes, K.; Lukacsovich, T.; Zhu, Y. Z.; Cattaneo, E.; Pandolfi, P. P.; Thompson, L. M.; Marsh, J. L. Science 2004, 304(5667), 100-104.

24. 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.; Ross, C. A. J Biol Chem 2009, 284(16), 10855-10867.

25. 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. Cell Cycle 2007, 6(23), 2970-2981.

26. 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.; Wetzel, R. Nat Struct Mol Biol 2009, 16(4), 380-389.

27. Bhattacharyya, A. M.; Thakur, A. K.; Wetzel, R. Proc Natl Acad Sci U S A 2005, 102(43), 15400-15405.

28. Chen, S. M.; Ferrone, F. A.; Wetzel, R. Proc Natl Acad Sci U S A 2002, 99(18), 11884-11889.

29. Kelley, N. W.; Huang, X.; Tam, S.; Spiess, C.; Frydman, J.; Pande, V. S. J Mol Biol 2009, 388(5), 919-927.

30. Humphrey, W.; Dalke, A.; Schulten, K. J Mol Graph 1996, 14(1), 33-38.

31. Sugita, Y.; Okamoto, Y. Chem Phys Lett 1999, 314(1-2), 141-151.

32. Vitalis, A.; Steffen, A.; Lyle, N.; Mao, A.; Pappu, R. V. J Chem Theory Comput 2009, *manuscript in preparation*.

33. Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; Van Gunsteren, W. F.; Mark, A. E. Angew Chem, Int Ed Engl 1999, 38(1-2), 236-240.

34. Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. J Chem Theory Comput 2008, 4(3), 435-447.

35. Tran, H. T.; Wang, X. L.; Pappu, R. V. Biochemistry 2005, 44(34), 11369-11380.

36. Vitalis, A.; Wang, X.; Pappu, R. V. Biophys J 2007, 93(6), 1923-1937.

37. Walgers, R.; Lee, T. C.; Cammers-Goodwin, A. J Am Chem Soc 1998, 120(20), 5073-5079.

38. Luo, P.; Baldwin, R. L. Biochemistry 1997, 36(27), 8413-8421.

39. Forood, B.; Feliciano, E. J.; Nambiar, K. P. Proc Natl Acad Sci U S A 1993, 90(3), 838-842.

40. Scholtz, J. M.; Qian, H.; York, E. J.; Stewart, J. M.; Baldwin, R. L. Biopolymers 1991, 31(13), 1463-1470.

41. Bernstein, S. L.; Wyttenbach, T.; Baumketner, A.; Shea, J. E.; Bitan, G.; Teplow, D. B.; Bowers, M. T. J Am Chem Soc 2005, 127(7), 2075-2084.

42. Marina, G. B.; Kirkitadze, D.; Lomakin, A.; Vollers, S. S.; Benedek, G. B.; Teplow, D. B. Proc Natl Acad Sci U S A 2003, 100(1), 330-335.

43. Robertson, A. L.; Horne, J.; Ellisdon, A. M.; Thomas, B.; Scanlon, M. J.; Bottomley, S. P. Biophys J 2008, 95(12), 5922-5930.

44. Nagai, Y.; Inui, T.; Popiel, H. A.; Fujikake, N.; Hasegawa, K.; Urade, Y.; Goto, Y.; Naiki, H.; Toda, T. Nat Struct Mol Biol 2007, 14(4), 332-340.

45. Walker, F. O. Lancet 2007, 369(9557), 218-228.

46. Lee, C. C.; Walters, R. H.; Murphy, R. M. Biochemistry 2007, 46(44), 12810-12820.

47. Takahashi, T.; Kikuchi, S.; Katada, S.; Nagai, Y.; Nishizawa, M.; Onodera, O. Hum Mol Genet 2008, 17(3), 345-356.

CHAPTER VII. CONCLUSIONS AND FUTURE WORK

The accomplishments of this thesis may be categorized into three broad areas. First, we have obtained a revised molecular picture of the biologically relevant process of the aggregation of polypeptides composed predominantly of glutamine. This is the major accomplishment and detailed in VII.3. Second, we have followed an interdisciplinary approach to our research: we demonstrated how a system of biomolecular interest represented at atomic resolution may be analyzed using ideas adopted from the conceptual physics of low complexity polymers. We summarize the relevance of this effort in VII.2. Third, we have created useful models and tools for other researchers in the process; specifically, we developed a novel continuum solvation model and a novel software package which both have unique features that should benefit the biomolecular simulation community. We describe those features and benefits next.

VII.1. Novel Methods for Computational Molecular Biophysics

VII.1.1. The ABSINTH Continuum Solvation Model

Chapter III presented the development and testing of a novel continuum solvation model[1] similar in paradigm to the EEF1 model of Lazaridis and Karplus.[2] The ABSINTH model (for self-**A**ssembly of **B**iomolecules **S**tudied by an **I**mplicit, **N**ovel, and **T**unable **H**amiltonian) is a major accomplishment of this thesis and consumed a considerable fraction of the time entering research. It has several noteworthy features which make it suitable for specific problems in the field of biomolecular simulations:

- ABSINTH is a computationally efficient implicit solvation model that nonetheless attempts to capture multi-body interactions. Typical MC calculations are slower by a factor of 3.0-5.0 than corresponding calculations using a gas-phase Hamiltonian. The situation is even better when equations of motion are integrated globally such as in MD: here, the factor is usually reduced to ~2.0. This places ABSINTH in a class with very simple implicit solvation models[2,3] and highly optimized variants of more complex methods.[4]
- The calibration of the model was somewhat unusual (see III.5 and III.6): we attempted to recapitulate the thermal stabilities and intrinsic disorder in biomolecular systems. We consciously avoided a parameterization with respect either to systems of little direct relevance (such as neat liquid and solution data of small molecules like activities, densities, heats of vaporization, etc.) or to *just* the stabilities of specific conformations of relevant biomolecules. Our parameterization paradigm predisposes ABSINTH to be used in studies of IDPs and in general of phenomena located near order-disorder transitions.
- ABSINTH inherently supports the simulation of co-solutes such as small molecules and counterions explicitly. This is because they are integrated into the same underlying framework. As is detailed below, this treatment may necessitate further improvements to the model in the future. Nonetheless, the ability to modulate solution conditions at this level of representation is an unusual feature for implicit solvation models.

One of the benefits of the ABSINTH paradigm is its flexibility. It opens several avenues for future improvements and extensions. First, pending the

availability of transfer free energies of solvation for the biomolecular model compounds (see Table 3.a) into organic solvents, we can conceive of simulations in other milieus. If the solvent is represented as a continuum, it may even be possible to simulate certain binary solvent mixtures that way. An example would be mixtures of water and urea as they are employed in CD spectroscopic experiments shown in Figure 6.4. Co-solvents are used in many applications in protein science, whether for chemical denaturation (urea, guanidinium chloride)[5] or for structural stabilization (TFE, THIP),[6] and reliable and exhaustive simulations of biomacromolecules under those conditions are currently lacking in the field.

The previous idea can be extended to address the simulation of lipid membrane environments in an implicit manner. There is precedent for this approach in general, and ABSINTH might offer a very natural framework to add to the list of descriptions currently available.[7-9] Strong interest in such methods has emerged in the biomolecular simulation field due to the increased focus on describing systems in environments more and more akin to what the native context might look like in the cell.

An application particularly suited to implicit solvation is the simulation of more appropriate ensembles for small volumes, *i.e.*, those in which the particle numbers fluctuate. The grand canonical ensemble is an underused description in biomolecular simulations,[10,11] primarily – or so we speculate – due to the inherent architectural constraints imposed by most simulation software packages. The accurate simulation of ensembles in which the pH remains constant[12] is a

specific application fitting into this broader framework and a direction for development in the immediate future. In the long run, a grand canonical description appears indispensable for the simulation of polyelectrolytes in the ABSINTH paradigm: this is another area in which our particular formulation of continuum solvation may excel compared to other implicit solvent models.

Lastly, there is one prominent weakness in the current treatment which needs to be addressed. The free energies of solvation of charged peptide and nucleic acid moieties have to be artificially lowered in order to prevent the formation of spurious salt bridges (see Table 3.a). This is fundamentally tied to the inability of the model to parse the system for the presence of multiple dielectric cavities. Given the polymeric context, which leads to some amount of desolvation, we need to be able to identify whether a second charged moiety (whether on a polymer or an ion in solution) shares the same dielectric cavity or not. There are several approaches to this rooted in either graph theory or simpler, geometric algorithms.[4]

In modern science, models gain traction by continued usage and by showing robust features in the process. Therefore, our primary aim for the immediate future should be to encourage the community to test the model on problems of interest to them and to continue to apply and evaluate the model ourselves. This way, the accuracy of our treatment can be continually refined by critical feedback received from a broad set of users.

VII.1.2. The CAMPARI Software Package

The development of a software package we termed CAMPARI (for **C**omputational **A**nalysis of **M**acromolecular **P**roperties **A**cross **R**esolutions and **I**nterfaces)[13] is an aspect of this work that has hardly been mentioned thus far. All the results in Chapters III-VI have been obtained using CAMPARI which was almost exclusively written by the author of this thesis.

The fundamental benefits and caveats of employing Monte Carlo methods in the conformational sampling of biomacromolecules have been reviewed recently.[14] The ability to "jump" in phase space is of fundamental appeal in rugged energy landscapes due to the inherent ability to cross barriers which might be associated with infeasible timescales when their crossing is simulated dynamically. Furthermore, the effective speed of conformational diffusion can be greatly enhanced in MC methods over an MD treatment if the landscapes are sufficiently flat. This is best illustrated using the simulations of macromolecular dimerization we carried out in Chapters IV, V, and VI: the simulation volumes are large (the droplet diameter typically was 40nm) and diffusion of the macromolecules would easily become prohibitively slow. A realistic modeling of the diffusional on-rate in those simulations would have been detrimental to our ability to quantify the thermodynamics of macromolecular association.

However, MC simulations are by no means the panacea of computational biophysics. Much like dynamics-based methods they have an application domain: that of systems with low average density and sufficient degeneracy. It is near-impossible for example to sample aqueous solutions of multiple, different

small molecules at ambient conditions. Here, density and specificity are both placed in a regime in which MC simulations become impractical. Conversely, our approach consisted of coupling the conformational equilibria to an implicit description of the solvent in which co-solutes present at low enough concentrations may be represented explicitly (see VII.1.1). It has been argued, not only by us[14] but by others as well,[15,16] that such a setting is amongst the scenarios conducive to MC methodologies exhibiting superior performance when compared to dynamics-based methods.

The CAMPARI software package has been developed as a simulation engine designed to take advantage of these putative benefits. It is intended for release in late summer of 2009 under a public license which allows the scientific community to benefit from our work at no cost. Its major features – and by extension its representation of the accomplishments of this thesis – are as follows:

- To our knowledge, it is the only simulation software presently available which offers comprehensive support for the ABSINTH implicit solvation model and the paradigm of the ABSINTH force field (see Chapter III and VII.1.1).
- CAMPARI offers a collection of move sets specifically designed to sample biomacromolecules when using standard molecular mechanics Hamiltonians. Support exists not only for polypeptides but also for polynucleotides.
- The software has validated support for most of the common molecular mechanics force fields, *i.e.*, CHARMM, AMBER, OPLS, and GROMOS. Since

CAMPARI also supports the basic paradigm underlying those force fields, comparative studies, a necessary endeavor whenever new simulation territory is explored, are feasible to an extent going beyond what is possible with other free simulation packages.

- Within the limits of our abilities and resources, CAMPARI is designed to be computationally efficient. Recent work on all major simulation software packages in this direction[17,18] indicates that researchers are less and less willing to spend their "experimental" resources, *i.e.*, CPU time, on inefficient algorithms.
- CAMPARI attempts to go beyond pure MC sampling and offers a variety of dynamics-based techniques, most prominently molecular and Langevin dynamics in both Cartesian and internal representations of the system. The ability to employ hybrid sampling protocols should carry substantial appeal in times when our ability to sample conformational space efficiently does not nearly grow as fast as our interest in more and more complex biological systems.
- Much of the analysis presented in Chapter II-VI is built into the software. Existing trajectory data may be re-analyzed using CAMPARI and informative quantities are computed on-the-fly for new simulations. This is a paradigm not typically encountered in simulation software but has proven a popular feature due to the significant time savings associated with not having to "manually" analyze all data *a posteriori*.

Future work on CAMPARI will undoubtedly be a collaborative effort. Several features may increase its relevance even more. Those include support for other biomolecules: lipids, polysaccharides, typical co-enzymes, ligands, and cellular metabolites, etc. Methodological advances are needed to be able to reliably simulate ensembles with constant pH or constant chemical potentials (see above). Continued efforts to make the software computationally more efficient are indispensable in ensuring that CAMPARI can become an accepted and widely employed tool in the biomolecular simulation community. Support for various platforms and architectures is an equally useful development direction to appeal to a wider audience of potential users. Lastly, we have actively considered the addition of alternative implicit solvent models to the available Hamiltonians, most prominently GB/SA-based models and the EEF1 model. This would allow comparative calculations across different continuum treatments of aqueous solvation with the ultimate goal of identifying consistent flaws in any of the models and cross-inform future improvements that way (see VII.1.1). It may sound surprising, but presently such efforts are hampered by the plethora of minute differences in implementation and interpretation of models one encounters from software package to software package.

VII.2. Interdisciplinary Aspects and the Role of Biophysicists

In I.1, we touched upon the interdisciplinary flavor an emergent field like biophysics holds due to its very definition: the application of rigorous and quantifiable methods, of concepts established in “pure” physics, and of a

corresponding school of thought to systems of biological interest. We pointed out the difficulty in remaining faithful to the details of the biological system given the inherently coarse-grained nature of the approach. Furthermore, we argued that we would continually reference our work to available experimental data. This second "interdisciplinary" aspect of connecting *in silico* to experimental data is discussed in this section as well.

The results in Chapter II illustrate the first aspect: we adopted order parameters and structural metrics from polymer physics to apply them to the problem of determining the conformational ensemble of polyglutamine.[19] Why was this possible, necessary, and beneficial at the same time? The first half of the answer lies in the low complexity of the primary sequence: polyglutamine is a homopolymer; therefore its properties may be predicted to be renormalizable albeit not in the rigorous sense of the term.[20] This allowed us to study a single chain length, yet make unequivocal predictions about chains of different lengths. This would of course not have been possible were the primary sequence that of a typical protein, *i.e.*, had we studied a non-random heteropolymer capable of adopting a specific fold. Polyglutamine's intrinsic disorder, which is well-characterized[21,22] and was confirmed multiple times in Chapters II-VI, provides the second half of the answer to the question above. The lack of a consistent preference for structural motifs allows the application of more coarse-grained metrics such as the ones presented in Chapter II: as an example, the scaling of internal distances with sequence separation (Figures 2.4 and 3.13 and Equation 2-6) may be considered. If the chains were to adopt a specific fold, this metric

would become uninformative: a plot like the one in Figure 2.4 would merely be a low-resolution transformation of the underlying structural motif.

However, it is not easy to convey the value of analyses like those presented in Chapter II to a broader audience. They appear less intuitive than atomistic pictures and condense information in a way that inevitably relies on averages over broad, disordered ensembles. One might argue that the reliance upon visualizations in the vein of structural biology has been an unfortunate bias in the biomolecular simulation field. We can speculate that the reason for this bias is that atomistic images of structures appeal to us because i) they go beyond the resolution provided by most experimental techniques; and ii) they satisfy our inherent conditioning and alignment of cognitive abilities toward visual culture.[23,24] The latter point may be illustrated by the concept of aesthetics often ascribed to "intuitive" visualizations, in particular in cell biology.[25] However, in the absence of X-ray crystal or NMR structures, *i.e.*, in the absence of directly comparable data, what is the intrinsic value of "representative" visualizations of simulation data?

The answer to this question is hierarchical. This thesis has attempted to circumvent the pitfall of referencing itself to *in silico* work as much as possible. Kratky profiles, for example, are direct readouts of SAXS experiments (see Figure 2.7).[26] Similarly, the scaling of internal distances can be probed experimentally (at least in theory) by a systematic FRET study with small enough dyes. By quantifying the scaling relationship of size and chain length, we obtained a single number from multiple sets of simulation data that is comparable

to the result of an FCS experiment (see Figure 3.12).[1,27] *If* we find that attempts to reproduce a set of experimental results as faithfully as possible are successful, we can *then* proceed to analyze our data further. We may identify microscopic driving forces within our models, derive mechanistic models, and create visual representations to illustrate our thinking. However, this is a meaningful and straightforward exercise only if our models are based on physical principles – whether implicitly or explicitly.

However, not all experimental techniques routinely employed in studies of biomolecular systems are understood well enough to be able to derive an efficient framework for their computation from simulation data. For example, the intrinsic fluorescence of tryptophan residues is a common readout of protein folding. Changes in macromolecular environment create a change in signal but the exact response function is poorly understood. NMR chemical shifts share the same feature, and their computation from simulation data is entirely empirical in nature.[28] In both cases, the theoretical framework of quantum dynamics would be needed to predict the experimental data *ab initio*.[29] With methods like these, coincidence of multiple, independent readouts is the most common way to address concerns about potential caveats underlying individual techniques. Figure 3.10 shows an analogous case for the interpretation of *in silico* data for the thermal melting of a small β-hairpin peptide: several coarse-grained, structural readouts are defined and shown to report on the same transition. Conversely, Figures 6.2, 6.4, and 6.5 present an example in which the lack of a

rigorous and quantitative interpretation of experimental data, in this case CD data, poses a problem for assessing the validity of our computational models.

Lastly, NMR methods present a particular challenge: often, multiple experimental constraints are obtained and fed into a computational algorithm which attempts to minimize deviations from the constraints given a chemically accurate representation of the system locally. This approach is prevalent in NMR structure determination and inherently assumes that all constraints need to be satisfied *simultaneously*. Consequently, several NMR "structures" of small peptides have been proposed.[30-32] It is very difficult to distinguish whether the experiments report on a canonical fold or on certain ensemble preferences sampled from an inherently disordered manifold. We addressed those concerns in detail in III.6.4. They point to the fact that presently there are cases in which a direct comparison between experimental and computational work is impossible.

What does the above discussion suggest with respect to the accomplishments of this thesis and with respect to the role of biophysicists in upcoming years? The first part is answered as follows: we have taken an unusual approach to quantifying the intrinsic properties of polyglutamine.[19,33,34] We were largely inspired by the physics of homopolymers and have attempted to add to the field of protein aggregation our insights and analyses.[35] Of course this line of thinking is not unique to us: *e.g.*, an influential paper by Kohn *et al*. demonstrates the application of low resolution experimental techniques to extract global characteristics inspired by polymer physics of a heterogeneous set of polypeptides.[36] Furthermore, our data and analyses directly suggest experiments

to be performed (*e.g.*, the systematic FRET analysis mentioned previously). In recent work, we have extended our approach to the biological problem of protamines.[37] Similar to the work in Chapter VI,[38] the realistic representation of the peptides has allowed us to remain faithful to the details of the system and to delineate global preferences from local heterogeneities which would be masked by analyses such as those presented in Chapters II and IV. An experimental test of some of the results of this study is underway (Crick and Pappu, unpublished data).

The second question aims at the role of biophysicists: the preceding discussion and the results of this thesis suggest that transfer of knowledge allows us to take a "fresh" look at biological problems and to arrive at novel and testable hypotheses. This is consistent with the very definition of being a biophysicist. Continuing efforts should be invested into the congruence of experimental and computational readouts, *i.e.*, into a better understanding of the concepts underlying common biophysical techniques and a better understanding of their limitations including guidance toward the design of suitable controls (see Chapter VI). Computational models should continue to be assessed carefully for their physical validity and qualitative and quantitative accuracy. Only then can mechanistic insights obtained from simulation data which extend beyond the resolution of experimental techniques live up to their potential. Lastly, communication skills will be a fundamental means in determining the success of biophysics.[39] We should always be able to apply our school of thought to a

problem domain in that problem domain's particular language and not expect others to translate for us.

VII.3. Aggregation of Polyglutamine and CAG Repeat Disease Pathogenesis

There is no cure for any of the nine exonic CAG repeat diseases known today. This statement was true a few years ago when research entering this thesis was started and remains true today upon its completion. So what have we accomplished from the point of view of the mission enabling us to conduct this research, that is, from the point of view of the mission of public health? Extensive discussions of the results of each chapter pertaining to this question are found in IV.4, V.4, and VI.4. We do not wish to repeat these in their entirety here but instead provide a brief evaluation of the results of this thesis in the context of understanding disease.

VII.3.1. Properties of Monomeric Polyglutamine and Implications for Disease

Recent work has placed pronounced emphasis on soluble species in delineating the causative agents for CAG repeat disease pathogenesis. This means that our understanding of the physicochemical and biological processes and mechanisms acting at the level of polyglutamine at very low concentrations and small copy numbers in cellular environments is crucial to understanding disease. A striking observation in CAG repeat diseases is the universally observed inverse correlation between disease age-of-onset and polyglutamine expansion length.[40] Consequently, research in Chapters II, IV, and V has

attempted to elucidate the structural preferences of monomeric polyglutamine homopolymers as a function of solvent quality and chain length. Our findings are as follows:

- At physiological temperatures and pressures, polyglutamine exists as compact, globular species in solution. Water is a poor solvent for polyglutamine and chain-chain interactions are preferred over forming an interface with the solvent.[19,34]
- Solvent quality may be modulated by a suitable control parameter such as temperature. We can induce a well-defined coil-to-globule transition as a function of solvent quality.[34] The very high θ-temperature of ~390K[33] we observe in an atomistic model using a realistic implicit solvation model known to reproduce melting temperatures for other polypeptides suggests that it may be difficult to adjust solvent quality for polyglutamine experimentally. This is indirectly supported by experimental results that show that peptides of the form GQ_NCK_2 with a cysteine-attached fluorophore remain collapsed and aggregation-prone even under chemically denaturing conditions at which most proteins unfold and are soluble (Crick and Pappu, unpublished data).
- The ensemble is disordered, in particular for longer chain lengths. No canonical secondary structure elements are detected in more than transient incarnations.[19,34] There is no sudden change in the conformational equilibria with chain length as was suggested repeatedly in the literature based on the threshold expansion length observed for HD ($N \approx 37$). We do not expect to ever

encounter critical chain length dependencies for the properties of the homopolymer, and this is consistent with the variable threshold lengths observed for the various CAG repeat diseases (see Table 1.a).[40]

- The conversion to a conformation rich in β-secondary structure is accompanied by a steep free energy barrier for polyglutamine.[33] Such a state would correspond to the putative arrangement in aggregates exhibiting amyloid-like characteristics.
- In poor solvent conditions, the interconversion between different disordered states for monomeric polyglutamine is hindered by the requirement to remain on the manifold of collapsed states. Conformational diffusion is slow, and the sluggish relaxation dynamics pose unique challenges for computer simulations of polyglutamine.[19]

The intrinsic disorder we observe for polyglutamine naturally places these peptides in the class of IDPs. Functional IDPs in the cell remain soluble by possessing large fractions of charged residues.[41,42] It may be assumed that they pose little to no stress for the cellular machinery concerned with protein folding and degradation with which they co-evolved. How is this possible? Molecular chaperones recognize mis- or unfolded proteins by exposed hydrophobic residues for which they provide large and non-specific binding surfaces.[43] Disordered sequences rich in charge are preferential binding motifs for the ubiquitin-proteasome degradation pathway and are efficiently degraded.[44] Conversely, well-folded proteins – for which water generally is a poor solvent – resist interactions with either chaperones or the proteasome by their micellar

architecture: hydrophobic and polar groups, including the polypeptide backbone, are sequestered away from the solvent by burial inside of a globular structure whose surface is coated with charged sidechains which interact favorable with the aqueous milieu encountered in the cell. Polyglutamine does not fit either of those two categories: water is a poor solvent but sequestration of solvophobic groups is by definition incomplete. Surface-exposed glutamine residues are expected to continually interact with molecular chaperones. The peptides may resist degradation via the ubiquitin-proteasome pathway due to their collapsed nature.[45-47]

A rigorous quantification of the conformational ensemble of homopolymeric polyglutamine has therefore enabled us to postulate a mechanism of toxicity given previous findings. However, the cellular relevance of hompolymeric polyglutamine is highly questionable. As is detailed in VII.3.2, we expect these peptides to fall out of solution very rapidly. It seems fairly unlikely that evolutionary pressure would allow cells to consistently waste energy on the creation and clearance of intracellular deposits even if we assume that the latter are non-toxic. The progressive development of visible aggregates in animal models of CAG repeat diseases suggests that polyQ-expanded proteins and their degradation products are largely kept soluble.[48] Given our results in Chapters IV and V, we cannot envision a scenario in which homopolymeric polyglutamine would remain soluble and be subjected to a controlled turnover cycle in the cell.

VII.3.2. Revised Aggregation Mechanisms for Polyglutamine

The results presented in Chapters II, IV, and V have led us to believe that the *intrinsic* properties of polyglutamine would give rise to an aggregation mechanism very different from what is proposed in the literature. It was touched upon several times that amorphous aggregates are not typically studied experimentally: they are morphologically ill-defined and are more likely viewed as waste products of failed protein preparations or experiments.[49] Moreover, the intrinsic heterogeneity renders a systematic characterization by experimental methods infeasible. Nonetheless, we conclude based upon the results in Chapters IV-VI that homopolymeric polyglutamine would rapidly aggregate to form an amorphous precipitate. Of course, observation of such a phenomenon in well-controlled conditions is very difficult due to the inability to chemically synthesize peptides without any flanking, charged residues added for solubility. Work in the 1960s had essentially established this on much larger polyamides.[50,51]

However, as the discussion in VII.3.1 highlights, we continue to view the intrinsic properties of polyglutamine as the predominant driving force in CAG repeat disease pathogenesis. The formation of soluble oligomers has been documented both *in vitro*[52] and *in vivo*,[53] and hence polyQ-mediated associations – whether with itself or with other aggregation-prone peptides present in the cellular milieu – remain a focal point of research in the field of exonic CAG repeat diseases. Recently, Bernacki and Murphy[54] argued that kinetic aggregation data using monomer loss as the experimental readout may be fit with a variety of

different mechanistic models including that brought forth by Wetzel and coworkers.[55] Distinguishing between models was shown to be difficult in the absence of more detailed data, in particular the complimentary readout of fibril numbers and sizes. Similarly, Morris *et al.*[56] demonstrated that a mechanism termed "the Finke-Watzky model of nucleation followed by autocatalytic surface growth" fits several independent sets of protein aggregation data reasonably well with only two free parameters.

It should be noted that in both of the aforementioned models no heterogeneities are included: neither in the (effective) nucleation nor in the elongation steps. Let us now consider a model in which monomers in solution quickly associate to give rise to a specific distribution of oligomers. If we consider monomer addition only, we have at equilibrium:

$$\begin{aligned} A + A &\rightleftharpoons A_2 \quad K_2 = \frac{[A_2]}{[A]^2} \\ A_i + A &\rightleftharpoons A_{i+1} \quad K_{i+1} = \frac{[A_{i+1}]}{[A_i][A]} \end{aligned} \tag{7-1}$$

In Equation 7-1, A is the aggregating (polymerizing) species, square brackets denote activities (from here on: concentrations), and the K_i are equilibrium constants. We can generate equations for the population of individual species, f_i, by observing conversation of mass:

$$\begin{aligned} f_i &= \frac{[A_i]}{c_t} = \frac{[A]}{c_t} \cdot \prod_{j=2}^{i} K_j [A] \\ c_t &= \sum_{i=1}^{\infty} i [A_i] \end{aligned} \tag{7-2}$$

Here, c_t is the total concentration of aggregating material in monomeric units. Equation 7-2 provides an implicit relationship between the concentration of free monomer and the total monomer concentration. Let us now consider the case where the infinite sum in Equation 7-2 is well-approximated by a finite sum; this is reasonable since we focus on *soluble* oligomers here. We define a maximum oligomer size $i_{\max}$ by setting K_i to be zero for $i>i_{\max}$. Equation 7-2 then reads (for clarity):

$$f_i = \begin{cases} \dfrac{[A]}{c_t} \cdot \prod_{j=2}^{i} K_j [A] & \text{for } i < i_{\max} \\ 0.0 & \text{else} \end{cases} \qquad (7\text{-}3)$$

$$c_t = \sum_{i=1}^{i_{\max}} i[A_i]$$

The graphical representations in Figures 6.9, 6.10 and 6.12 led us to believe that this may be a reasonable scenario for the sequence constructs of interest to us here. Let us now assume that there is a very slow and irreversible conversion of large enough oligomers A_i with $i \geq i_{min}$ to fibrillar species:

$$\begin{aligned} &A_{i_{\min}} \xrightarrow{k_{i_{\min}}} F_{i_{\min}} \\ &A_{i_{\min}+1} \xrightarrow{k_{i_{\min}+1}} F_{i_{\min}+1} \\ &\ldots \\ &A_{i_{\max}} \xrightarrow{k_{i_{\max}}} F_{i_{\max}} \end{aligned} \qquad (7\text{-}4)$$

In Equation 7-4 F_i denotes a fibrillar species composed of i monomers. They are formed through a unimolecular process which we can assume to correspond to an internal re-arrangement akin to the ideas presented in Figure 5.16. Due to the slowness of the process we may safely assume that the oligomer distribution re-

equilibrates rapidly (pre-equilibrium assumption). We can then write a rate equation for *heterogeneous* fibril formation as:

$$\begin{aligned}\frac{d[F_i]}{dt} &= k_i[A_i] = k_i[A]\cdot\prod_{j=2}^{i} K_j[A] \\ \frac{dc_p}{dt} &= \sum_{i=i_{\min}}^{i_{\max}} k_i[A_i] = \sum_{i=i_{\min}}^{i_{\max}} k_i[A]\cdot\prod_{j=2}^{i} K_j[A]\end{aligned} \tag{7-5}$$

In Equation 7-5, c_p is the total concentration of growing ends (independent of the size of the particular fibril nucleus) and the k_i are a set of heterogeneous nucleation rate constants. Consequently, we can treat elongation via monomer addition as:

$$\begin{aligned}&F_i + A \underset{k_-}{\overset{k_+}{\rightleftharpoons}} F_{i+1} \\ &\frac{dc_f}{dt} = k_+ c_p[A] - k_- c_p \\ &\frac{d^2 c_f}{dt^2} = k_+ c_p \frac{d[A]}{dt} + \left(k_+[A] - k_-\right)\cdot\sum_{i=i_{\min}}^{i_{\max}} k_i[A]\cdot\prod_{j=2}^{i} K_j[A]\end{aligned} \tag{7-6}$$

Here, k_+ is the elongation rate constant for monomer addition, k_- is the rate constant for monomer loss from a growing fibril, and c_f is the total concentration of monomers incorporated into fibrils.

If we focus on the initial rate of aggregation and treat fibril elongation as irreversible and homogeneous, we may assume, at least for certain values of the K_i, that the free monomer concentration [A] is static and determined by its pre-equilibrium value in the absence of any fibrillar aggregates. This concentration a_0 is obtained implicitly via Equation 7-3. Then, we can integrate Equation 7-6 similar to Equation 1-6:

$$c_f(t) = \frac{1}{2} k_+ t^2 a_0 \cdot \sum_{i=i_{\min}}^{i_{\max}} k_i a_0 \cdot \prod_{j=2}^{i} K_j a_0 \qquad (7\text{-}7)$$

This form allows us to now pursue an analysis identical to the one proposed by Wetzel and colleagues. Figure 1.2 has shown that the analysis in itself is fairly robust given a homogeneously nucleated process. But what happens in the case of heterogeneous nucleation to the concentration dependence of the initial rate of aggregation? This question is answered directly by Equation 7-7. First, we see that the slope of a double logarithmic plot still convolutes elongation and nucleation processes much like in the original analysis. Second, we recognize that the effective dependence on total concentration is now not at all guaranteed to yield a well-defined integer slope in a double logarithmic plot since it crucially depends on the relationship of a_0 and c_t, which is given by Equation 7-3. It is in parts the *significant* population of off-pathway, *i.e.*, fibril-incompetent oligomers and in parts the heterogeneity of the nucleation mechanism itself that fundamentally alters the concentration dependence. Interestingly, Equation 7-7 recovers the model postulated by Wetzel (Equation 1-7) if both i_{min} and i_{max} adopt a value of two. The only difference is that the pre-equilibrium constant K^{n^*} becomes a pre-equilibrium dimerization constant K_2 and that the bimolecular nucleus elongation rate constant k^*_+ becomes a unimolecular nucleation rate constant k_2. The model proposed here is therefore a valid generalization of the homogeneous nucleation model.

We illustrate the altered concentration dependencies and resultant nucleus size estimates in Figure 7.1. We define a simple model for both the K_i

and the k_i by drawing them from a normal distribution with well-defined means and variances. We can then determine large numbers of apparent slopes from a double logarithmic plot analogous to Panel D of Figure 1.2 to study the impact of heterogeneous nucleation on estimates of the nucleus size assuming homogeneous nucleation, *i.e.*, n^* following the idea of Equation 1-8:

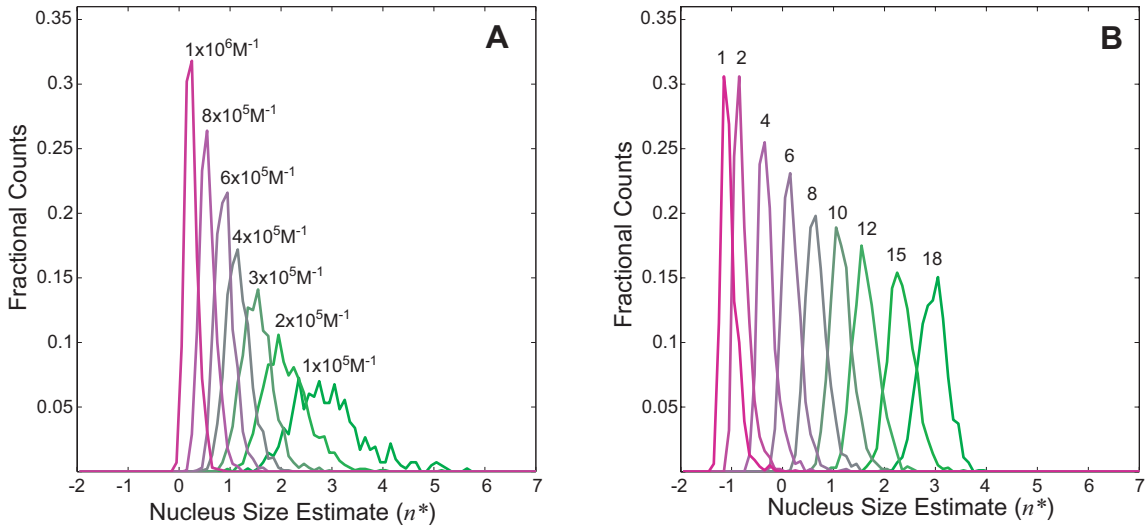

Figure 7.1: Apparent nucleus size estimates n^* for simulated aggregation data for a heterogeneously nucleated process. The parameters of the distribution to generate the k_i were constant throughout (mean and standard deviation are $10^{-2}s^{-1}$; values are adjusted to zero if drawn as negative), and i_{max} was always 20. Panel A shows histograms of apparent nucleus size estimates for a minimum fibril-forming oligomer size of ten and for different values of the mean of the distribution of K_i. The variance was fixed and set to $1x10^5M^{-1}$. Again, values were adjusted to zero if drawn as negative. Cases with bad parameter values were discarded. Every histogram represents 10^3 simulated datasets. Within each dataset, 450 free monomer concentrations up to the high μM-range were used to obtain species distributions and total concentrations via

Equation 7-3. Linear regression of a double logarithmic plot of the constant terms in Equation 7-7 against c_t yielded n^*+2 according to Equation 1-8. To mimic experimental conditions, only values for c_t of 1-500μM were considered for the line fit. Panel B shows an analogous plot. Here, i_{min} is varied and the mean of the K_i is fixed to $4 \times 10^5 M^{-1}$. The obtained nucleus sizes are somewhat concentration range-dependent (not shown). This convolution of effects is another important point to consider for example in the work of Thakur *et al*.[57]

Panels A and B of Figure 7.1 show that the estimated nucleus size from a heterogeneously nucleated process (as defined above) depends strongly on which regime the system is prepared in. If most of the aggregating material exists in fibril-incompetent forms (small K_i and/or large i_{min}), the apparent nucleus size is large but smaller than i_{min}. Fractional values are easily observed similar to experimental data.[22,57] If, however, large fractions of the aggregating material are present in soluble oligomers large enough to promote fibril formation (large K_i and/or small i_{min}), the limiting value for the apparent nucleus size appears to be close to zero (corresponding to a slope of 2.0) but can in fact be less. This finding is inadvertently consistent with the autocatalytic surface growth mechanism proposed by Morris *et al.*, whose initial rate dependence on total concentration would similarly yield a slope of 2.0.[56] If we interpret the analysis according to Equation 1-8, nucleus size estimates of fractional numbers smaller than unity are entirely possible (the apparent n^* can indeed be negative as reported by Thakur *et al.*). It should be noted that the parameters chosen for Figure 7.1 are arbitrary.

The model merely illustrates how experimental data may be more reasonably explained by a fundamentally different process than the one postulated.

We can therefore propose a revised aggregation mechanism for glutamine-rich, heteropolymeric polypeptides. Taken together, the results of Chapters IV, V, and VI and of Figure 7.1 are very much consistent with a mechanism outlined in steps (g), (h), (i), and (j) of Figure 5.16. Under typical experimental conditions *in vitro*, polyQ-expanded peptides spontaneously form soluble oligomers.[52] The properties of these species are controlled by the location and prevalence of solubilizing (charged) amino acid residues.[38] Large enough oligomers may provide a water-deprived environment for chains on the inside of soluble oligomers. A rate-limiting conversion to a fibrillar species occurs and induces phase separation which leads to a quantifiable readout of monomer loss.[33,54] Figure 7.1 suggests that the experimental work of Wetzel and coworkers is carried out in conditions where soluble oligomers are prevalent. This provides a clean explanation for the fractional values of less than unity observed for (homogeneous) nuclei sizes experimentally.[22,55,57]

VII.3.3. Therapeutic Strategies

The lack of successful and general therapeutic strategies means that present-day treatment of patients is entirely symptomatic at the level of the behavioral and physiological phenotypes (see I.2.2).[58] However, various molecular therapeutic strategies have been discussed and begun to be tested in cellular or animal models of disease.[59] Many of those ideas hinge upon the

availability of gene therapy, *i.e.*, the effective, cost-efficient, well-controlled delivery of exogenous genes to the affected tissues. The actual molecular therapeutic is then made available by the cellular machinery and underlies normal metabolic processes. As an inherently endogenous material, such polypeptide or RNA drugs circumvent several side effects possible with small molecule drugs due for example to metabolic by-products. The expression of exogenous antibodies recognizing mutant huntingtin ("intrabodies") has been demonstrated to exhibit neuroprotective effects in *Drosophila* and mouse models of disease.[60,61] However, with current medicinal technology, a transfer of this strategy is not yet feasible. If it were, efforts to re-engineer those intrabodies to work as polyglutamine-specific proteases would seem like an extremely valuable research target. Since molecular medicine has not matured to this stage yet, the remainder of this section discusses the relevance of our findings in the context of more feasible strategies.

Modulation of the PQCS

In 2009, a considerable body of literature exists that supports the idea that an up-regulation of the activity of molecular chaperones may have a cytoprotective effect.[62-68] The underlying idea is consistent with the dominant hypothesis of toxicity outlined in I.2.4: glutamine-rich peptides including soluble forms and – to a lesser extent – precipitated forms cause stress to the PQCS which the cell eventually succumbs to. Our results are very much consistent with this hypothesis as was detailed in VI.5 and VII.3.1.

But how does this represent a route to therapy? The only current and feasible application being pursued is the administration of small molecules known to change the protein levels of those species involved in the PQCS. Compounds such as geldanamycin and derivatives have shown potential in fly models.[64] The advantage of this strategy is that it might be generalizable to other protein aggregation diseases as a universal involvement of the PQCS has been proposed.[69] However, this thesis does not shed any light on the molecular mechanisms underlying the efficacy of such compounds; hence, any further discussion is omitted.

Small Molecule Inhibitors of Aggregation

Our results suggest that an efficient sequestration of reactive interfaces mediating deleterious protein-protein interactions may represent an effective strategy to prevent the PQCS from being impaired by the presence of glutamine-rich, disordered peptides. They also suggest that the unique challenge posed by polyglutamine might lie in its ability to remain soluble enough to form liquid-like monomers and oligomers in solution which are amenable to reversible associations. What about homopolymeric sequences that are expected to be even more aggregation-prone? Evidence suggests that evolutionary pressure has rigorously prevented the presence of gene expansions which would give rise to peptide fragments rich in hydrophobic residues.[70]

Various medium- or high-throughput screening assays have been set up in recent years to identify compounds – among those with some ability to cross the blood-brain barrier – that can inhibit aggregation of polyQ-expanded protein

in vitro or *in vivo*.[71,72] One of the problems imposed by the absence of a detailed mechanistic understanding of the process is the selection of evaluation criteria. For example, our data suggest that the formation of fibrillar aggregates need not be an informative readout of therapeutic potential (see VII.3.2). Not surprisingly, results pertaining to the cytoprotective effects of the amyloid-binding and aggregation-inhibiting dye Congo Red remain controversial.[73,74] Instead, we propose that an *in vitro* characterization of oligomer distributions and aggregation rates for sequence constructs identified by analysis of proteolytic fragments may yield a much more informative assay for the screening of small molecules. However, severe concerns remain: i) the polyQ-expanded protein is constantly being produced by the cell; ii) it is very unlikely that a small molecule identified in a screen has high enough *selectivity* to not interfere with other cellular – in particular self-assembly – processes when applied in suitable dosages, and iii) it is well-known that screening assays may yield generic false positives referred to as chemical aggregators.[75] Compounds in that latter class could easily prove toxic due to their reported ability to sequester functional proteins from the surrounding milieu and to (at least partially) unfold them.[76]

Structural Drug Design

We concluded in Chapters II, IV, and V that there is no consensus structural motif present in homopolymeric polyglutamine at the level of monomers and dimers. We speculate that the same will hold true for larger, soluble oligomers. The early stages of polyglutamine aggregation would be intrinsically disordered in terms of their protein secondary, tertiary, and quaternary structures.

Even if we stipulate that β-secondary structure is a common motif in glutamine-rich aggregates, that motif would by no means be unique. Our results strongly suggest that structure-based design targeting the polyglutamine segment is not a viable therapeutic strategy – primarily due to the absence of a consensus motif and due to the low concentration of glutamine-rich fragments *in vivo*.

Conversely, Chapter VI elucidated that wild-type flanking sequences may very well exhibit structural preferences. Our results argue that those preferences might be transient in nature and that their prevalence will depend on the relative lengths of the structured motif and the polyQ-expansion. Ironically, it might be easier to engineer structurally designed drugs that recognize fragments carrying non-pathogenic polyQ-expansions than to engineer drugs recognizing those with longer polyglutamine stretches. Presently, polypeptides which specifically bind polyQ-expanded protein have been identified via screening[77] or as antibodies[78] and not via targeted design. Recently, the "exposed β-sheet hypothesis" has been brought forth[59] which was formulated primarily based upon results obtained for a thioredoxin fusion protein[79] as discussed in I.2.7. We argue that this is misleading as a general hypothesis: our studies of the intrinsic properties of polyglutamine attribute little significance to β-secondary structure. The two lines of thought are easily reconciled if we neglect the secondary structure component and propose a modified "exposed reactive polyglutamine hypothesis" as outlined above.[80]

Summary

In summary, our findings as a whole question our ability to *selectively* interfere with polyglutamine-mediated aggregation processes in cellular environments. The lack of selectivity suggests that many compounds will also be cytotoxic and may not be suitable drug candidates for the treatment of exonic CAG repeat diseases. The cellular machinery which successfully handles the stress imposed by the mutant proteins is already in place. Assisting the PQCS may represent a viable strategy for the treatment of all protein misfolding and aggregation diseases. Beyond that, our results indicate that future research should focus on identifying naturally occurring, glutamine-rich fragments to be able to screen them for putative drug targets within the flanking sequences. Research along the lines of the work presented in Chapter VI will be indispensable in providing a molecular characterization of such polypeptides.

VII.3.4 Future Directions

Our work has remained limited by the system sizes we are able to study. Even though we obtained an atomistic picture of the dimerization of polyglutamine at various chain lengths and sequence contexts, all our results on higher-order assemblies remain speculative in nature. Concepts adopted from basic polymer physics have assisted us in deriving meaningful predictions from our data (see Chapters II, IV, and V). However, Chapter VI has elucidated that our intuition may only guide us to a certain point.

Future *in silico* work should therefore strive to bridge the gap between different length- and timescales by further coarse-graining the representation of

the system. This is a difficult objective if our interest is to continue to capture the underlying physics of the assembly process, in particular for heteropolymeric polypeptide sequences. Work has begun which will address the stability of amyloid-like aggregated phases (Lyle, Vitalis, and Pappu, unpublished). Ultimately, a demonstration of the phenomenon implied in Equation 7-4, *i.e.*, the structural re-arrangement of a disordered oligomer into an ordered fibril is an ambitious but worthwhile goal for future computational research.

Furthermore, we should strive to identify alternative experimental readouts which report on the sizes and numbers of soluble oligomers. A better understanding of metrics employed routinely in the protein aggregation field – such as ThT binding or CD spectroscopy – appears as a highly desirable goal for future work as well. Then, the predictions outlined in VII.3.1 and VII.3.2 become testable by experimental techniques – a universal goal of all *in silico* work.

Lastly, our focus should migrate beyond the realm of CAG repeat diseases. While detrimental for the families involved, prevalence is generally low. Other, more prevalent neurodegenerative diseases, in particular Alzheimer's, share basic mechanistic features which make them attractive targets to apply our methodology, thinking, and resources to. Work in this direction has already begun (Ramasubramanian and Pappu, unpublished).

VII.4. Bibliography

1. Vitalis, A.; Pappu, R. V. J Comput Chem 2009, 30(5), 673-699.
2. Lazaridis, T.; Karplus, M. Prot Struct Funct Gen 1999, 35(2), 133-152.

3. Ferrara, P.; Apostolakis, J.; Caflisch, A. Prot Struct Funct Gen 2002, 46(1), 24-33.

4. Haberthur, U.; Caflisch, A. J Comput Chem 2008, 29(5), 701-715.

5. Hua, L.; Zhou, R.; Thirumalai C, D.; Berne, B. J. Proc Natl Acad Sci U S A 2008, 105(44), 16928-16933.

6. Luo, P.; Baldwin, R. L. Biochemistry 1997, 36(27), 8413-8421.

7. Lazaridis, T. Prot Struct Funct Gen 2003, 52(2), 176-192.

8. Ulmschneider, M. B.; Ulmschneider, J. P.; Sansom, M. S. P.; Di Nola, A. Biophys J 2007, 92(7), 2338-2349.

9. Grossfield, A. In Computational Modeling of Membranes; Feller, S., Ed., 2008, p 131-157.

10. Vitalis, A.; Baker, N. A.; McCammon, J. A. Mol Simul 2004, 30(1), 45-61.

11. Im, W.; Roux, B. J Chem Phys 2001, 115(10), 4850-4861.

12. Mongan, J.; Case, D. A. Curr Opin Struct Biol 2005, 15(2), 157-163.

13. Vitalis, A.; Steffen, A.; Lyle, N.; Mao, A.; Pappu, R. V. J Chem Theory Comput 2009, *manuscript in preparation*.

14. Vitalis, A.; Pappu, R. V. In Annual Reports in Computational Chemistry; Baker, N. A., Ed.: *in press*, 2009.

15. Ulmschneider, J. P.; Ulmschneider, M. B.; Di Nola, A. J Phys Chem B 2006, 110(33), 16733-16742.

16. Jorgensen, W. L.; Tirado-Rives, J. J Comput Chem 2005, 26(16), 1689-1700.

17. Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. Journal of Chemical Theory and Computation 2008, 4(3), 435-447.

18. Case, D. A.; Darden, T. A.; T.E. Cheatham, I.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; M. Crowley; Walker, R. C.; Zhang, W.; Merz, K. M.; B.Wang; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; K.F.Wong; Paesani, F.; Vanicek, J.; X.Wu;

Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. University of California, San Francisco 2008.

19. Vitalis, A.; Wang, X.; Pappu, R. V. Biophys J 2007, 93(6), 1923-1937.

20. Schäfer, L. Excluded Volume Effects in Polymer Solutions as Explained by the Renormalization Group; Springer: Berlin, 1999.

21. Masino, L.; Kelly, G.; Leonard, K.; Trottier, Y.; Pastore, A. FEBS Lett 2002, 513(2-3), 267-272.

22. Chen, S. M.; Ferrone, F. A.; Wetzel, R. Proc Natl Acad Sci U S A 2002, 99(18), 11884-11889.

23. Keirns, C. J Hist Biol 1999, 32(1), 163-195.

24. Gooding, D. C. J Cogn Cult 2004, 4(3-4), 551-593.

25. Wingate, R.; Kwint, M. Nat Rev Neurosci 2006, 7(9), 745-752.

26. Glatter, O.; Kratky, O. Small Angle X-Ray Scattering; Academic Press: London, 1982.

27. Crick, S. L.; Jayaraman, M.; Frieden, C.; Wetzel, R.; Pappu, R. V. Proc Natl Acad Sci U S A 2006, 103(45), 16764-16769.

28. Neal, S.; Nip, A. M.; Zhang, H. Y.; Wishart, D. S. J Biomol NMR 2003, 26(3), 215-240.

29. Barone, V.; Cimino, P.; Crescenzi, O.; Pavone, M. J Mol Struct Theochem 2007, 811(1-3), 323-335.

30. Kortemme, T.; Ramirez-Alvarado, M.; Serrano, L. Science 1998, 281(5374), 253-256.

31. Blanco, F. J.; Rivas, G.; Serrano, L. Nat Struct Biol 1994, 1(9), 584-590.

32. Cochran, A. G.; Skelton, N. J.; Starovasnik, M. A. Proc Natl Acad Sci U S A 2001, 98(10), 5578-5583.

33. Vitalis, A.; Lyle, N.; Pappu, R. V. Biophys J 2009, *in press*.

34. Vitalis, A.; Wang, X.; Pappu, R. V. J Mol Biol 2008, 384(1), 279-297.

35. Pappu, R. V.; Wang, X.; Vitalis, A.; Crick, S. L. Arch Biochem Biophys 2007, 469(1), 132-141.

36. 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.; Plaxco, K. W. Proc Natl Acad Sci U S A 2004, 101(34), 12491-12496.

37. Mao, A.; Vitalis, A.; Pappu, R. V. Proc Natl Acad Sci U S A 2009, *to be submitted*.

38. Williamson, T. E.; Vitalis, A.; Crick, S. L.; Pappu, R. V. Nat Struct Mol Biol 2009, *to be submitted*.

39. Gooding, D. C.; Addis, T. R. Found Sci 2008, 13(1), 17.

40. Walker, F. O. Lancet 2007, 369(9557), 218-228.

41. Romero, P.; Obradovic, Z.; Li, X.; Garner, E. C.; Brown, C. J.; Dunker, A. K. Prot Struct Funct Gen 2001, 42(1), 38-48.

42. Uversky, V. N.; Gillespie, J. R.; Fink, A. L. Prot Struct Funct Gen 2000, 41(3), 415-427.

43. Liberek, K.; Lewandowska, A.; Zietkiewicz, S. EMBO J 2008, 27(2), 328-335.

44. Prakash, S.; Tian, L.; Ratliff, K. S.; Lehotzky, R. E.; Matouschek, A. Nat Struct Mol Biol 2004, 11(9), 830-837.

45. Wang, J. J.; Wang, C. E.; Orr, A.; Tydlacka, S.; Li, S. H.; Li, X. J. J Cell Biol 2008, 180(6), 1177-1189.

46. Bennett, E. J.; Bence, N. F.; Jayakumar, R.; Kopito, R. R. Mol Cell 2005, 17(3), 351-365.

47. Venkatraman, P.; Wetzel, R.; Tanaka, M.; Nukina, N.; Goldberg, A. L. Mol Cell 2004, 14(1), 95-104.

48. Mangiarini, L.; Sathasivam, K.; Seller, M.; Cozens, B.; Harper, A.; Hetherington, C.; Lawton, M.; Trottier, Y.; Lehrach, H.; Davies, S. W.; Bates, G. P. Cell 1996, 87(3), 493-506.

49. Morris, A. M.; Watzky, M. A.; Finke, R. G. Biochim Biophys Acta 2009, 1794(3), 375-397.

50. Krull, L. H.; Wall, J. S. Biochemistry 1966, 5(5), 1521-1527.

51. Krull, L. H.; Wall, J. S.; Zobel, H.; Dimler, R. J. Biochemistry 1965, 4(4), 626-633.

52. Lee, C. C.; Walters, R. H.; Murphy, R. M. Biochemistry 2007, 46(44), 12810-12820.

53. Takahashi, T.; Kikuchi, S.; Katada, S.; Nagai, Y.; Nishizawa, M.; Onodera, O. Hum Mol Genet 2008, 17(3), 345-356.

54. Bernacki, J. P.; Murphy, R. M. Biophys J 2009, 96(7), 2871-2887.

55. Bhattacharyya, A. M.; Thakur, A. K.; Wetzel, R. Proc Natl Acad Sci U S A 2005, 102(43), 15400-15405.

56. Morris, A. M.; Watzky, M. A.; Agar, J. N.; Finke, R. G. Biochemistry 2008, 47(8), 2413-2427.

57. 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.; Wetzel, R. Nat Struct Mol Biol 2009, 16(4), 380-389.

58. Ross, C. A.; Margolis, R. L.; Rosenblatt, A.; Ranen, N. G.; Becher, M. W.; Aylward, E. Medicine 1997, 76(5), 305-338.

59. Nagai, Y.; Popiel, H. A. Curr Pharm Des 2008, 14(30), 3267-3279.

60. Wang, C. E.; Zhou, H.; McGuire, J. R.; Cerullo, V.; Lee, B.; Li, S. H.; Li, X. J. J Cell Biol 2008, 181(5), 803-816.

61. Wolfgang, W. J.; Miller, T. W.; Webster, J. M.; Huston, J. S.; Thompson, L. M.; Marsh, J. L.; Messer, A. Proc Natl Acad Sci U S A 2005, 102(32), 11563-11568.

62. Parfitt, D. A.; Michael, G. J.; Vermeulen, E. G. M.; Prodromou, N. V.; Webb, T. R.; Gallo, J. M.; Cheetham, M. E.; Nicoll, W. S.; Blatch, G. L.; Chapple, J. P. Hum Mol Genet 2009, 18(9), 1556-1565.

63. Raychaudhuri, S.; Sinha, M.; Mukhopadhyay, D.; Bhattacharyya, N. P. Hum Mol Genet 2008, 17(2), 240-255.

64. Fujikake, N.; Nagai, Y.; Popiel, H. A.; Okamoto, Y.; Yamaguchi, M.; Toda, T. J Biol Chem 2008, 283(38), 26188-26197.

65. Rujano, M. A.; Kampinga, H. H.; Salomons, F. A. Exp Cell Res 2007, 313(16), 3568.

66. Cummings, C. J.; Sun, Y. L.; Opal, P.; Antalffy, B.; Mestril, R.; Orr, H. T.; Dillmann, W. H.; Zoghbi, H. Y. Hum Mol Genet 2001, 10(14), 1511-1518.

67. Warrick, J. M.; Chan, H. Y. E.; Gray-Board, G. L.; Chai, Y. H.; Paulson, H. L.; Bonini, N. M. Nat Genet 1999, 23(4), 425-428.

68. McLear, J. A.; Lebrecht, D.; Messer, A.; Wolfgang, W. J. FASEB J 2008, 22(6), 2003-2011.

69. Muchowski, P. J.; Wacker, J. L. Nat Rev Neurosci 2005, 6(1), 11-22.

70. Dorsman, J. C.; Pepers, B.; Langenberg, D.; Kerkdijk, H.; Ijszenga, M.; Den Dunnen, J. T.; Roos, R. A. C.; Van Ommen, G. J. B. Hum Mol Genet 2002, 11(13), 1487-1496.

71. Ehrnhoefer, D. E.; Duennwald, M.; Markovic, P.; Wacker, J. L.; Engemann, S.; Roark, M.; Legleiter, J.; Marsh, J. L.; Thompson, L. M.; Lindquist, S.; Muchowski, P. J.; Wanker, E. E. Hum Mol Genet 2006, 15(18), 2743-2751.

72. Zhang, X. Q.; Smith, D. L.; Merlin, A. B.; Engemann, S.; Russel, D. E.; Roark, M.; Washington, S. L.; Maxwell, M. M.; Marsh, J. L.; Thompson, L. M.; Wanker, E. E.; Young, A. B.; Housman, D. E.; Bates, G. P.; Sherman, M. Y.; Kazantsev, A. G. Proc Natl Acad Sci U S A 2005, 102(3), 892-897.

73. Heiser, V.; Scherzinger, E.; Boeddrich, A.; Nordhoff, E.; Lurz, R.; Schugardt, N.; Lehrach, H.; Wanker, E. E. Proc Natl Acad Sci U S A 2000, 97(12), 6739-6744.

74. Wood, N. I.; Pallier, P. N.; Wanderer, J.; Morton, A. J. Neurobiol Dis 2007, 25(2), 342-353.

75. McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. J Med Chem 2002, 45(8), 1712-1722.

76. Coan, K. E. D.; Maltby, D. A.; Burlingame, A. L.; Shoichet, B. K. J Med Chem 2009, 52(7), 2067-2075.

77. Nagai, Y.; Tucker, T.; Ren, H.; Kenan, D. J.; Henderson, B. S.; Keene, J. D.; Strittmatter, W. J.; Burke, J. R. J Biol Chem 2000, 275(14), 10437-10442.

78. Kayed, R.; Head, E.; Thompson, J. L.; McIntire, T. M.; Milton, S. C.; Cotman, C. W.; Glabe, C. G. Science 2003, 300(5618), 486-489.

79. Nagai, Y.; Inui, T.; Popiel, H. A.; Fujikake, N.; Hasegawa, K.; Urade, Y.; Goto, Y.; Naiki, H.; Toda, T. Nat Struct Mol Biol 2007, 14(4), 332-340.

80. Chen, S.; Berthelier, V.; Yang, W.; Wetzel, R. J Mol Biol 2001, 311(1), 173-182.