Intrinsically Disordered Proteins: Mechanics, Assemblies, and Structural Transitions

by

Mehran Bagheri

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
For the Ph.D. degree in Physics

Ottawa-Carleton Institute of Physics
Department of Physics, Faculty of Science
University of Ottawa
Ottawa, Canada

uOttawa

Dedicated to my loving family. To my parents, Mahin and Mohammad whose deep devotion and sacrifice enable me to follow my dreams; To my brother, Majid, for his endless encouragement in furtherance of my dreams. To my beautiful wife, Niloofar, whose continual inspiration and support keep my dreams alive and eventually turn them to reality.

Abstract

Proteins are essential parts of living organisms that initiate and control almost all cellular processes. Despite the widely accepted belief that all functional proteins fold into stable and well-defined three-dimensional (3D) structures mandatory for protein activity, the existence of biologically functional disordered proteins has been increasingly recognized during past two decades. Proteins with inherent structural disorder, commonly known as intrinsically disordered proteins (IDPs), play many roles in a biological context. However, in contrast to their folded counterparts, they are dynamically unstructured and typically fluctuate among many conformations even while performing biological functions. In fact, it is this dynamical structural heterogeneity that that allows for IDPs to interact with other biological macromolecules in unique ways. Moreover, while a majority of proteins in eukaryotic proteomes have been found to have intrinsically disordered regions (IDR), the mechanisms by which protein disorder gives rise to biological functionality is still not well understood. Through a series of simulation studies on specific systems, this thesis probes several aspects of the emerging structure-function paradigm of IDPs, namely the mechanics, intermolecular assembly, and structural transitions occurring in these proteins.

The lack of well-defined 3D structure in IDPs gives rise to distinct mechanical properties, the subject of the first study in the thesis on the elasticity of a elastomeric gluten-mimetic polypeptide with an intrinsically disordered character. This disordered polypeptide was shown to exhibit distinctively variable elastic response to a wide range of tensions, which a classical worm-like chain model failed to accurately describe, thus requiring a molecular-level analysis.

IDPs are frequently involved in protein-protein interactions, the focus of the second study on the propensity of an IDR, the B domain in dynamin-related protein 1 (Dpr1), to self-assemble into dimer structures while remaining disordered in all solution conditions. Despite a hypothesized auto-inhibitory role for this domain in Dpr1 that was assumed to be triggered by an disordered-to-order transition, the B domains in solution showed no tendency

to form ordered structures even in the presence of order promoting osmolytes. Instead, self-association in the presence of osmolyte was found to occur by favorable intermolecular intereactions between specific region on the surface of the B-domains.

Other IDPs do undergo a disorder-to-order transition in response to environmental cues, in ways that are unique disordered proteins, the focus of the last study on intermolecular ordering transitions in silk-like proteins. Factors such as protein sequence and physical tension were investigated, and results suggested that tyrosine residues in the key silk sequence motifs promote templating of beta structure from disordered precursors and that elongational stresses preferentialy stabilize antiparallel beta-sheet order. Together, these three computational studies provide insight into the nature of the structure-function mechanisms of IDPs.

Acknowledgements

First, and foremost, I would like to express my sincere gratitude to my advisor Dr. James L. Harden for the opportunity to work in his laboratory. It was a great privilege to have worked and learned under the mentorship of such a knowledgeable and ambitious scientist with a great personality. Thank you for showing me nothing but patience and unparalleled professionalism over several years.

I would like to thank senior scientist Dr. Ebrahim Karimi, Dr. Andrew Pelling, and Dr. Bela Joos, for continual inspiration and stimulating conversations during my graduate study. I also thank all my kind-hearted and caring colleague in the physics department, particularly Rahil Golipour and Kristina Haase. Thank you to the welcoming staff of the physics department, namely Madeleine Thomas and Lisa Murphy.

Last but not least, thank you to my beautiful wife who are the source of my everlasting drive for success and without whom I could not have made it this far.

Thank you, everyone.

Statement of Originality

To the best of my knowledge, I hereby declare that this thesis establish original work generated during my Ph.D. studies under the supervision of Dr. James L. Harden, in the Centre for Interdisciplinary Nanophysics laboratories in the Department of Physics at the University of Ottawa. Chapter 4 was published whereas Chapter 2,3 and Appendix A,B are prepared for peer-reviewed scientific journals.

I solely generated all simulations data and perform statistical analysis to develop computational results while Dr. James L. Harden contributed to the direction and editing of all manuscripts. Other researchers also contributed to this work performing experiments were acknowledged in each chapter. Dr. Ammon Posey and Dr. Blake Hill introduced B domain and TMAO osmolytes explain in Chapter 3. Lin Yinan and Benjamin Partlow, who were Ph.D. Candidates at the time under the supervision of Dr. David Kaplan, provided all experimental results and contributed to writing the experimental portion of the manuscript in Chapter 4 and Appendix A, respectively. Python package (Appendix B) was written and developed by me; while Emilie Fortin, undergraduate summer student, edited the code and scripted a graphical user interface for the package.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction: Functional Protein Disorder

1.1 Principles of Protein Structure and Function

Proteins are biomolecules composed of amino acid residues. They are essential parts of living organisms that initiate and control almost all cellular processes, including DNA replication, signal transduction, cell adhesion, cell division, transporting other molecules, and many others. For many years, there was a strong belief that a protein amino acid sequence drives its transition to a particular well-defined three-dimensional (3D) equilibrium structure which determines protein function. This model is referred to as the *sequence-structure-function* paradigm; the transition process is called *folding* and final well-ordered structure is referred to as the *native* or *folded* structure. Such a native structure is almost fully compact and typically has a globule shape for water-soluble proteins, and are thereby classified as *globular* proteins. However, it has recently been discovered that many globular proteins without a well-defined 3D structure have important biological functions. There are numerous biologically functional proteins that are entirely disordered, the so-called intrinsically disordered proteins (IDPs). More commonly, there are proteins that possess partial 3D order and have intrinsically disordered regions (IDRs) that play a crucial role in their function. The structure of these disordered regions and their ability to accomplish their functions, have been the subject of many recent investigations. However, the description of the unfolded state resembling IDPs/IDRs and their functions require a comprehensive and more general understanding of protein structure and function mechanisms. This chapter introduces generic knowledge and terminology that has been employed for many years to explain the structure and function of globular ordered proteins and how it can be extended to the disordered proteins.

1.1.1 Amino acid sequence as protein signature

Proteins contain one or more linear chain, also known as a *polypeptide*, consisting of subunits called amino acids, which are covalently linked together via peptide bonds. Amino acids are organic molecules which generally contain a central carbon atom covalently bound to a carboxylic acid group and an amino group to form the so-called *protein backbone*. The central carbon, known as alpha-carbon (C_α), also is bonded to a hydrogen and another group of atoms, often denoted as the *sidechain* group. Each amino acid has a unique sidechain group. A schematic view of a polypeptide chain, indicating the peptide bond and chemical structure of its amino acid residue is given in Figure 1.1.

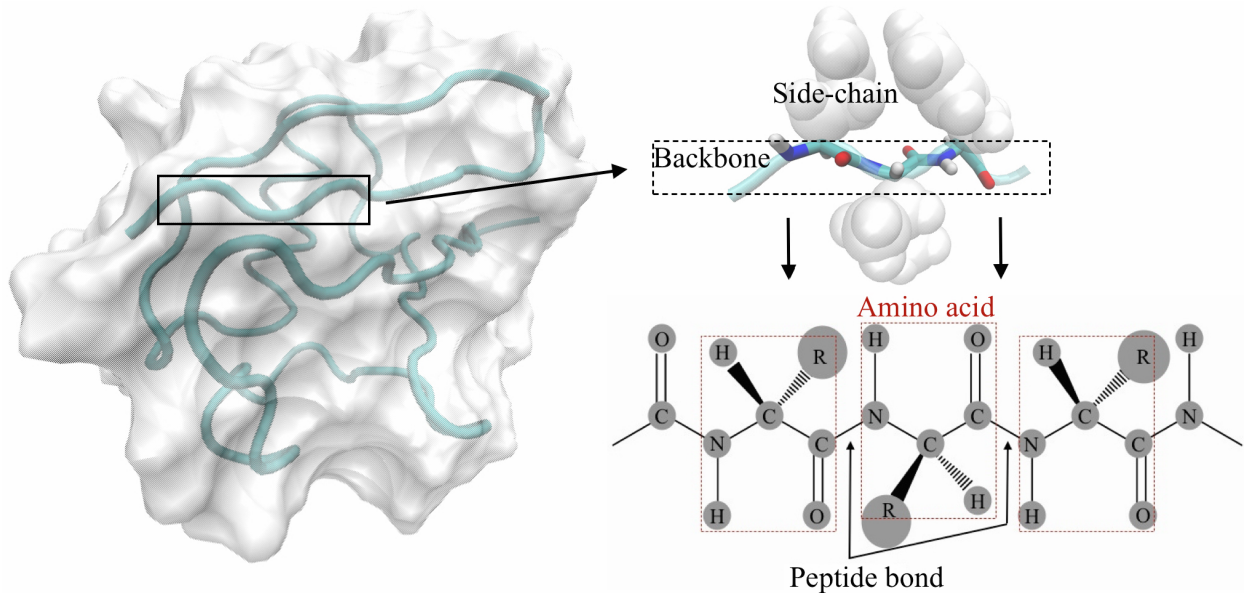


Figure 1.1: Schematic diagram showing a protein structures and its atomic constituents. The geometric shape of a protein is defined by its sidechain groups, coloured in white, and its backbone coloured in cyan. The peptide bond and amino acid chemical structure are also indicated, wherein sidechain groups are noted as R. The protein structure represented is that of ubiquitin (PDB ID:1UBG).

There are 20 amino acids that share a common backbone [†], but various sidechain groups which determine physicochemical characteristics (e.g. acidic, basic, polar, hydrophobic, small) of the amino acids. Intermolecular interactions between amino acid residues along the chain and surrounding solvent molecules determine the coordinates of the equilibrium positions of all atoms of the protein. These interactions can be modulated by the linear arrangement of amino acids in the polypeptide chain, called protein *sequence*, that ultimately drive a polypeptide chain to its native conformation (if it has one). This was initially proposed by Christian B. Anfinsen, based on his research on the enzyme ribonuclease folding mechanism [1], the work that has been awarded the Nobel Prize in 1972. He postulated that protein native structure is exclusively determined by its amino acid sequence and other external factors (e.g. temperature and solvent conditions), leading to a stable, unique structure (if these factors do not disrupt the folding process).

The native structure of a protein is governed by the specific weak interactions between amino acid residues, for which there is no creation or disruption of covalent bonds during the

[†]The only exception is proline which has its sidechain covalently linked to its backbone amino nitrogen

folding process. The most common interaction between atoms is the van der Waals interaction, which occurs between permanent/induced dipoles [2]. Also, sidechain groups of amino acids may possess a net electric charge, resulting in strong electrostatic interactions or it may interact favorably or unfavorably with water molecules depending on its hydrophobicity [3]. While amino acids sidechains are dominantly involved in these interactions, the primary interaction in the protein folding process is hydrogen bonds which are usually formed between backbone carbonyl oxygens and amide hydrogens and contribute significantly to protein backbone shape and final native conformation.

1.1.2 Description of protein structure

Describing the native structure of a protein based on the coordinates of all of its atoms to characterize its structure and function is not only inconvenient, but is also not necessary. Therefore, a hierarchical description has traditionally been used to present the different levels of protein structure. It is generally agreed that the amino acid sequence of a polypeptide determines its equilibrium conformation; thus, at the most basic, informational level a protein can be described simply by its sequence, known as *primary structure*, although it does not correspond to any physical structure. Backbone hydrogen bonds induce *local* polypeptide chain conformations, termed *secondary structure*, which are characterized by the spatial arrangement of the chain backbone and by the hydrogen bond pattern.

The most common secondary structure elements are α-helices and β-sheets, the latter of which can be composed of parallel or antiparallel β-strands. A schematic view of these elements is given in Figure 1.2. Note that Linus Pauling, who also received a Nobel prize for his contribution into the nature of the chemical bond, had predicted the existence of these local structures [4], which were confirmed a few years later when protein structure was solved for the first time using x-ray diffraction [5], in the 1950s.

Next, these local structures within a single chain interact via the amino acid sidechain groups and may fold into the unique stable conformation called *tertiary structure*. Regions of a chain unit, also so-called *domain*, may then interact via their tertiary structure surface and possibly form a stable oligomeric species termed a *quaternary structure*. Tertiary and quaternary structures are described by the type of secondary structure which is dominant in their content and contact surface, respectively. For instance, “globin fold” is a class of tertiary structure where there are eight helices, but no other major structural feature and

myoglobin has the globin tertiary structure.

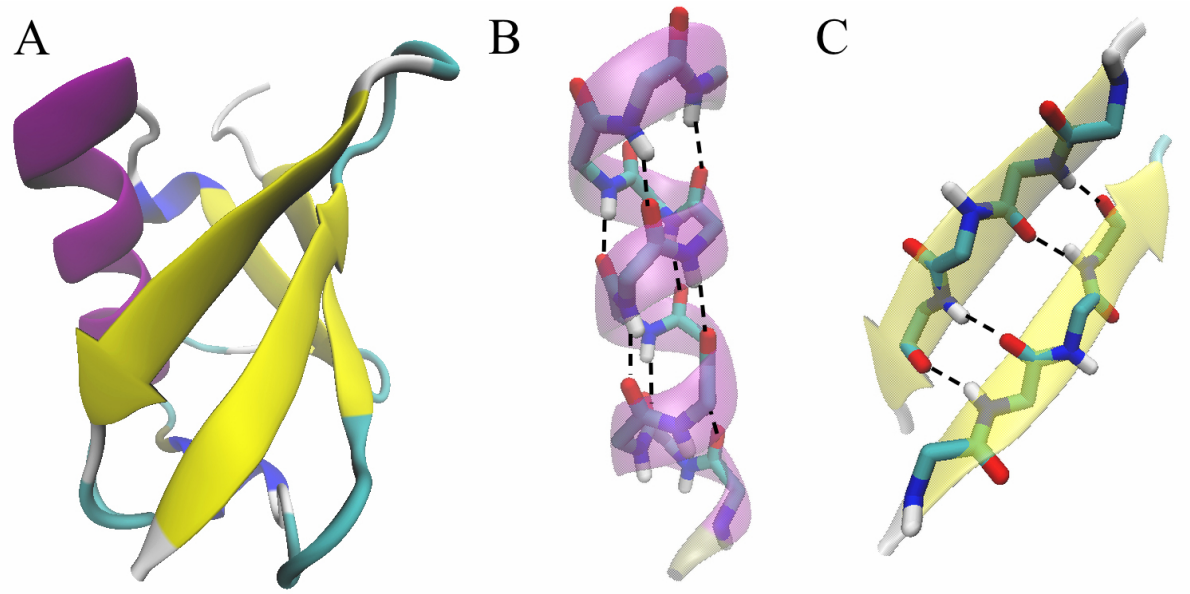


Figure 1.2: Conventional member of protein secondary structures. A) Schematic diagram of ubiquitin protein (PDB:1UBQ) indicating its α-helix (purple ribbon) and β-sheet (yellow arrows) structures. These conformations have a repetitive hydrogen bond pattern, in the polypeptide backbone, between hydrogen atoms in the amide group and oxygen atoms in the carbonyl groups. Hydrogen bonds are depicted in dashed black lines for α-helix (B) and β-sheet (C) structures along the backbone atoms (oxygen, red; nitrogen, blue; carbon, cyan; hydrogen, white)

A protein secondary structure can be described in terms of torsion angles around backbone bonds. However, the four atoms of the peptide bonds (O=C—N–H) mainly reside in a plane due to the partial double-bond between the carbon and the nitrogen [6]. Thus, no rotation is possible around that C—N bond and only two torsion angles around the bond on either side of the C_α, the so-called *dihedral angle*, are required to describe the overall spatial structure of a local polypeptide conformation. Traditionally, the dihedral angle around C^{i-1}–N^i—C^i_α–C^i bond is denoted by "ϕ", and around N^i–C^i_α—C^i–N^{i+1} bond is denoted by "ψ" (See Figure 1.3). The distribution of these angles for a given protein is typically reported in form of a two-dimensional (ϕ,ψ) plot known as a *Ramachandran plot* [7]. Individual secondary structure elements occur most favorably at the specific regions in the Ramachandran plot according to their energetically preferred dihedral angles. For instance α-helix dihedral angles (ϕ, ψ) are around (-60°, -45°) and beta strands angles are near (-130°,135°).

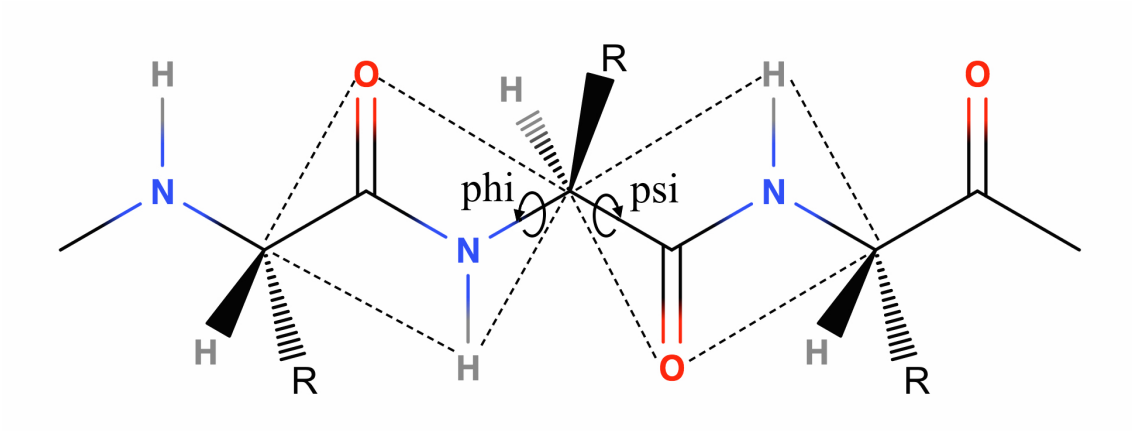


Figure 1.3: Dihedral angles around alpha-carbon (C_α) and planar peptide bonds. Four atoms of the peptide bonds, on either side of C_α stay within a plane shown in dashed rectangular region. Thus, protein backbone conformation can be described by the two torsion angles around the bond on either side of the C_α. A useful descriptor of local conformation of a polypeptide chain is the pair of dihedral angles phi (ϕ) and psi (ψ), which are torsion angles around N^i—C^i and C^i_α—C^i, respectively.

1.1.3 Folding kinetics and thermodynamics

The protein sequence is the main driving force for the folding of a protein into the native structure. Reversibility of the folding process implies that the native structure is thermodynamically stable and corresponding to the global minimum in the protein free energy landscape [8, 9]. Other than the unique native conformation, there are an astronomically large number of possible conformations even for a small protein with about 100 amino acid residues. Consequently, discovery of the native structure through a random sampling of conformations would take hundred years which is significantly longer than the typical biological folding timescale, which ranges from microsecond to a few hours at most. This contradiction, known as *Levinthal paradox* [10], suggests the existence of a systematic process for folding of the protein, the so-called *folding pathway*. This implies that the polypeptide amino acid sequence not only determines the native conformation (thermodynamic aspect of folding) but also the folding pathway leading the chain to its native state (kinetics aspect of folding).

It has been generally accepted that the folding pathway contains multiple short-lived intermediate states which are biased to evolving toward the native state [11]. Thus proteins have a free energy landscape resembling a funnel, where protein conformations are directed toward a possibly unique native state at the bottom of the funnel-like free-energy

landscape [12]. The folding pathway often involves a continuous transition from unfolded to native structure. However, a simplified but useful approach is to describe the conformational of a protein using a two state model, where the protein is either in an unfolded state, also called *denatured* state, or in the folded (native) state.

In contrast, partially unfolded intermediates state between the native and the denatured states have been experimentally observed using guanidine solutions or in pH variation for numerous, but not all, proteins [13,14]. These proteins exhibited one or more distinguishable intermediate states that contain similar secondary structure to the native structure but are not as compact as folded state. Alternatively, there are some proteins initially and rapidly form a disordered, compact conformation, arising from clustering of hydrophobic sidechains and electrostatic interactions between charged groups, which then may transform to the native structure though a complex reorganization of the core into well-packed secondary structures [15]. In this rather disordered intermediate conformation, also called *molten globule* (MG), sidechain groups have enhanced local freedom of motion relative to the native state in which they are tightly packed, but significant barriers to larger scale conformational changes. Overall, the folding pathways can divided into two or three different types,

$$\text{Unfolded/denatured} \rightleftharpoons \text{native}$$

or

$$\text{Unfolded/denatured} \rightleftharpoons \text{molten globule} \rightleftharpoons \text{native}$$

The folding process involves a large number of inter- and intramolecular interactions occurring at different length scales. Local interactions such as hydrogen bonds induce secondary structures whereas long-range interactions between protein sidechain groups, such as hydrophobic contact, play an essential role in tightly packing the secondary structures and shaping the tertiary structure of the native state. These interactions may also induce multiple barriers along the folding pathway causing the free-energy landscape to be rough and contain multiple intermediate states. These intermediate states, which define the kinetics of the folding process, can be transient and lead to the native structure or be a meta-stable structure such as molten globule. Moreover, intermediate states are usually not as compact as native states, in spite of having similar secondary structure or even partial tertiary structure. Size/compactness of protein states and corresponding intramolecular interactions are commonly described using polymer models, which are explained extensively in background section in Chapter 2, and can serve as an indication of the state of a protein in solution.

1.2 Appearance of Disordered Protein

A long time before Anfisntion and other scientists postulated the sequence-structure paradigm in the late 90s, Fischer in 1894 proposed the *lock-and-key* model for proteins with catalytic activity [16]. In his model, an enzyme and its substrate need to "fit to each other like a lock and key in order to exert a chemical effect on each other". Thus, it was postulated that a unique well-defined 3D structure is required to provide a tight fit between the binding pocket and substrate, in order to accelerate the reaction by decreasing the energy barrier. It was subsequently verified experimentally that proteins in denatured states lack their enzyme activity [17]. Although it was first proposed for enzymes, it has been shown that well-defined 3D structure is a prerequisite for many other types of protein function [18, 19]. This idea currently is known as *structure-function* paradigm and it is in agreement with many well-defined (at atomic resolution) 3D structures available in the Protein Data Bank (PDB) [20], many of which were obtained using X-ray crystallography technique.

The first evidence that the lock-and-key model is not universally valid was reported by Karush [21], who noticed that serum albumin protein has almost the same binding affinity for small molecules with a totally different shapes. In other words, serum albumin protein is able to adopt different 3D structures to fit substrates with different shapes. Thus, a unique, rigid, folded structure is not strictly necessary for proteins to engage in specific inter-molecular association processes. Based on Karush's observations, a modified version of the structure-function paradigm, the *induced fit* model, was proposed [22]. In this model, a given protein binding site may have multiple well-defined 3D conformations with similar energies where its ultimate conformation depends on its binding partner. In this picture, a protein adopts a particular structure from its conformational ensembles that is the best spatial complementary structure for a given binding partner. The induced fit model was further supported by the notable difference between the crystallized structure of an enzyme, A isozyme of yeast hexokinase, in presence and absence of an associated substrate, glucose [23], indicated the occurrence of conformational transition for the protein upon binding to its binding partner.

The above structure-function models are consistent with the prerequisite of well-defined 3D structure for protein functionality, wherein a protein may have a unique rigid binding pocket (lock-and-key), or adapt from its conformational ensembles upon interacting with a binding partner (induced-fit). However, there is also ample evidence of unstructured regions of proteins, which often appeared as missing coordinates in X-ray crystallography experi-

ments. In these experiments, missing coordinates of atoms could be due to the failure to solve the phase problem, crystal defects, or alternatively be due to mobile atoms which unable to scatter X-rays coherently. In the early stages of protein X-ray crystallography in the 1960s, researchers assumed that these missing regions were not important and they often removed these disordered segments to enable crystallization. The structure of these regions with missing electron density were not discovered until pioneering nuclear magnetic resonance (NMR) experiments indicated that the tail of histone H5 has a dynamic and non-compact shape [24]. Subsequently, many proteins which are partially or fully disordered and yet have known functionality were studied using NMR [25, 26].

Due to their unstructured characteristics, disordered proteins are usually not as compact as the native state of ordered proteins. Analogous to the induced-fit model, disordered proteins may have a rough free energy landscape where the folded state (with the global minimum energy) is only accessible in specific physicochemical environments, such as through binding to specific substrates. On the other hand, some of these proteins do not even have a well-defined 3D structure with a global minimum energy and may constantly be transformed between different non-equilibrium conformations in a manner similar to the molten globular state in folding pathway of globular proteins. The lack of a well-defined ground state 3D structure is an essential privilege of disordered proteins which enables them to perform some biological functions exclusively (See Section 1.3 for more details). On the other hand, this unstructured feature of intrinsically disordered proteins (IDPs) needs to be encoded in their amino acid sequences, which define protein intramolecular interactions.

In comparison with foldable proteins, the amino acid composition of disordered proteins is noticeably enriched in lysine, glutamine, serine, glutamic acid and proline, while depleted in cysteine, tryptophan, isoleucine, tyrosine, phenylalanine, leucine, histidine, valine, asparagines and methionine amino acids [27]. The former group of amino acids are the so-called *disorder-promoting* amino acids, while the latter are termed *order-promoting*. This is clearly shown in Figure 1.4, where frequencies of amino acids in disordered proteins are compared individually with their occurrence in ordered proteins.

As expected, polar and charged residues are considered as disorder-promoting, perhaps due to their disruptive effect on protein folding into compact structures, for instance via electrostatic repulsion in the core of protein or their tendency to favour direct interact with solvent molecules. In contrast, the hydrophobicity of most of the order-promoting amino acid facilitates a pathway to a compact structure required for folding. In addition to amino

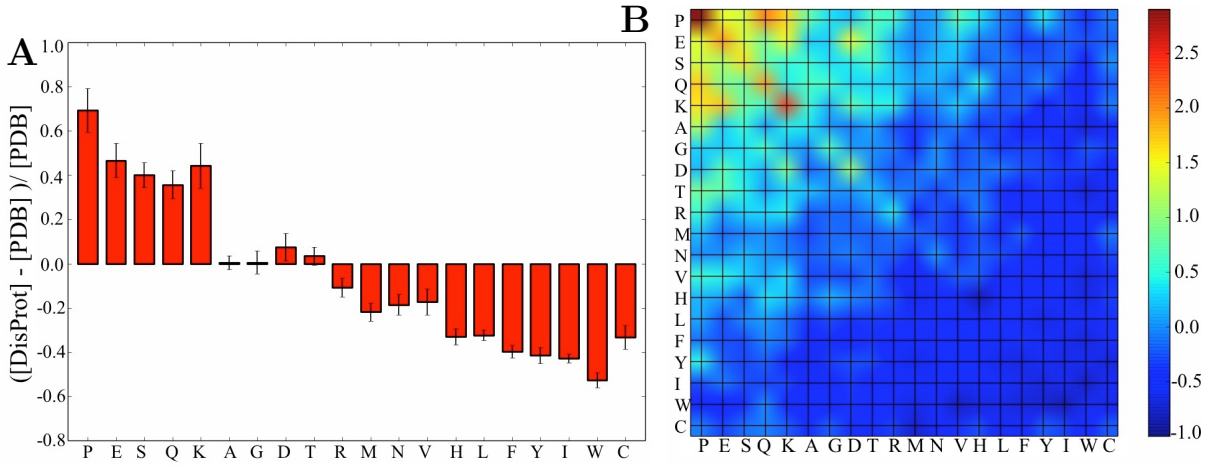


Figure 1.4: Propensity for individual amino acids and pairs of amino acids in disordered proteins. (A) Fractional difference in the amino acid frequency between disordered datasets (DisProt database [28] and ordered datasets proteins (PDB bank [20]) measured for each amino acid residue. [PDB] and [Disprot] indicate the content of a given amino acid in the ordered and disordered proteins database, respectively. Disordered proteins have a tendency to be enriched in residues with positive bars, whereas negative bars correspond depleted residues from these proteins. Amino acid types were ranked according to their decreasing disorder-promoting potential. (B) a similar analysis as in part A performed for all possible pairs of amino acids, where the most the frequent amino acid in disordered protein datasets was indicated in dark red color. Non-redundant protein sequences selected from proteins with less than %25 sequence identity (as measured using x-ray crystallography with resolution $<$ 3.5 Å) and posted before 2013 in the PDB bank [20] were used for ordered data sets; and sequences in the Disprot v6.02 [28] were used for the disordered proteins datasets. Statistical results were calculated using the bootstrapping method adopted from [29].

acid composition, the arrangement of amino acids in a sequence is expected to play a role in the protein structure. In particular, the frequency of amino acid pairs are also significantly different between ordered and disordered proteins. This is because dihedral angles play an essential role in determining polypeptide conformation, and identical amino acids can possess different dihedral angles depending on their adjacent residues. Indeed, an amino acid residue adjacent to bulky amino acids such as tyrosine has limited range of accessible dihedral angles compared with the same residue adjacent to the small amino acid such as glycine. Also, two like-charged amino acids are not energetically favourable in close proximity. Results of amino acid pairs analysis is shown in Figure 1.4 B, indicating distinct dipeptide sequences found in disordered proteins.

Moreover, investigation of the amino acid sequences of known intrinsically disordered proteins indicates statistically significant local patterns of amino acids or their properties such as charge, polarity, and size [30]. For example, silk and glutamine proteins studied in this work contain highly repetitive motif (template of short amino acid composition) in their disordered regions. As a result, IDP/IDRs contain noticeably lower sequence complexity with respect to ordered globular proteins [31].

1.3 Disorder Advantageous in Function

Intrinsically disordered proteins are newly discovered members of the protein kingdom [32], which are involved in numerous cellular processes. In fact, nearly half of the eukaryotic proteome is predicted to contain disordered regions [33]. This family of proteins also is implicated in various diseases, such as Alzheimer's [34] and cancers [35]. The lack of well-defined 3D structure, in favour of dynamically sampling a range of conformation from fully extended polypeptide chains to collapsed globules, make them well suited for regulatory and signaling processes [36], where specific response to an external element is required. IDPs/IDRs are able to possess different conformational states and their functions mainly arise from this structural heterogeneity. In the absence of a binding partner, an IDP/IDR may continually adopt different structures from its conformational ensembles and behave as an entropic chain. On the other hand, IDPs/IDRs frequently undergo a disorder-to-order transition in a biological process upon binding to another molecule, a process known as *coupled folding* [37]. In addition, they may also remain in disordered conformations while they are binding to other proteins, thereby forming a so-called *fuzzy complex* [38] to perform their biological function [39, 40]. Therefore, IDP/IDRs play their biological roles essentially in different conformational modes that are explained in more detail in the following subsections.

1.3.1 Structural ensemble and entropic chain

Some IDP/IDRs are able to carry out biological functions without undergoing a disorder-to-order transition. The dynamic structure of disordered proteins/regions enables them to fluctuate between a large number of conformations as a polypeptide chain with coil-liked characteristics. These flexible regions enable proteins to search for binding partners smoothly by providing enhanced spatial separation of functional regions in the manner of an *Entropic*

Spacers/Bristle [41] and thereby avoid undesirable association or aggregation. They also may act as *Linkers* that increase binding specificity and affinity by structural rearrangement and alternative binding interactions along their binding partner [42].

On the other hand, in absence of a binding partner, fluctuations between a large number of conformational states provide mechanical properties analogous to random-coil. By stretching or compressing a disordered region, it loses conformational freedom and the conformational entropy of the protein is reduced. As a result, the chain responds reversibly to restore its high entropy. A great example of such molecular spring is Titin, a gigantic protein with the size of $\sim 0.5\ \mu m$, that is responsible for the passive elasticity of muscles. Titin includes long disordered regions rich in Proline(P), Glutamic acids (E), Valine (V), and Lysine (K), known as PEVK domains. This disordered region behaves as a rubber-like entropic spring, which acts to restore stretched muscle cells to their relaxed length [43]. The elastic properties of such PEVK domains have been found to approximately obey the worm-like chain model in force-extension measurements [44]. There are disordered regions in numerous proteins, such Elastin, Resilin, and Wheat-Gluten, that consist of repetitive sequence motifs which impart such elastic properties to the proteins. In Chapter 2, the mechanics of a biomimetic Wheat-Gluten polypeptide were studied at the molecular level.

1.3.2 Disorder interaction and regulation of protein complex

Disordered proteins are frequently involved in protein-protein interaction networks [45], where functional disorder enables them to bind to multiple partners used for assembly and facilitates regulation of the different segments of the complex assemblies [46]. In these assemblies, disordered regions serve as a molecular glue that can be activated by modification of some global parameters of the protein environment such as solution pH, ionic strength, and interactions with membranes and ligands [47]. In this context, an IDP/IDPR may form a fuzzy complex, in which it remains in a disordered conformation while participating in protein assemblies [48, 49], or it may undergo coupled-folding and adopted an ordered structure, such as in the SNARE protein that is responsible for fusion of vesicle into the target membrane in eukaryotes [50].

Membrane fusion is the process whereby two distinct lipid membranes merge to into a single contiguous one. This is the mechanism which vesicles employ in order to deposit their contents (e.g., hormones, receptors, transporters, channels or adhesion molecules) into the

target membrane. The SNARE complex is responsible for bringing membranes into close proximity and regulating membrane fusion for vesicles in neurons. The SNARE complex consists of syntaxin-1, SNAP-25 (synaptosome-associated protein of 25 kDa) and VAMP (vesicle-associated membrane protein), where the first two are included in the target plasma membrane and the last one is positioned on the vesicle membrane. The SNARE proteins contain large disordered regions that spontaneously assemble into a remarkably stable four-helix bundle that zippers up. As results, they force two membranes into close proximity and mediate membrane fusion [51]. Drp1 is another protein associated with lipid membrane regulation which has a disordered domain, known as the B domain, with possible role of protein-membrane interactions. The structure and function of the B domain investigated in Chapter 3.

Structural transitions in intrinsically disordered protein assemblies is an exceptional phenomena that is still not well understood. While structural changes induced by binding-partner proteins can be quite complicated, transitions can also be induced by changes in physicochemical parameters in the protein's environment, particularly pH and temperature. Such physicochemical-induced disorder-to-order transitions have common features with those that occur with protein species that have a well-defined folded state. Moreover, such transitions also occur in many protein-systems, where the describe large-scale ordered structures that assemble from disordered proteins that may also have common features with the case of quaternary structure formation in folded protein systems. In Chapter 4, one such disorder-to-order transition involving the association of silk-mimetic polypeptides was investigated wherein silk protein undergoes pH reduction and folds into beta-sheet structure, while there is no noticeable charge residue in the motif involved in beta crystallines.

1.4 Thesis overview

Unstructured Intrinsically disordered proteins are distinct from ordered proteins in sequence, structure, and function. IDPs/IDRs are involved in various vital functions in cells in both disordered or induced folded states, yet little is know about how they function and their possible privilege over ordered proteins which requires a more detailed approach to determine. Studying disordered proteins experimentally is challenging due to their conformational fluctuation which occur in short time scales. In contrast, computer simulation methods, providing atomistic resolution to the protein structure and folding mechanisms, are able to

directly probe highly dynamic conformations and thus have emerged as powerful tools to characterize disordered regions in this family of proteins. In my thesis work, I used atomistic molecular dynamics (MD) simulations to study several intrinsically disordered systems, with each study focusing on a different physical aspect of these systems. Several of these studies were directly motivated by experimental work conducted in the Harden lab, University of Ottawa, or in the labs of external collaborators.

In the first investigation (Chapter 2), the mechanical behaviour of a disordered polypeptide was studied and compared to simple polymer models of extension under applied tension. The polypeptide design was based on a repetitive motif of the *gluten* protein family, a well-known *elastomeric* protein. This polypeptide, which has a strictly repetitive sequence (and therefore has low sequence complexity), forms a relatively compact, disordered conformation in force-free, steady-state conditions. Its extensional response to applied tension was found to exhibit a number of different regimes depending on the magnitude of the applied tension, as described in detail in the text, and thus was found to be more complex than simple coil-like or semi-flexible polymer systems. The nature of the polypeptide response in each force regime are discussed in the context of the fundamental mechanical properties of IDPs/IDRs. This work were first presented at the March 2014 meeting of American Physics Society (APS) and a manuscript is currently being prepared for publication in a peer-reviewed journal.

The next study (Chapter 3) was on the conformation and self-interaction of the disordered B domain of the Drp1 protein involved in fission of membranes. This work was motivated by our experimental collaborators, Dr. Ammon Posey and Prof. Blake Hill, who found that the absence of the B domain decreases Drp1 efficiency and furthermore that B domain was able to form assemblies in solution conditions which induce folding in many disordered structures. The structure and self-assembling pathway (coupled-folding or fuzzy complex) of the B domain were investigated to understand their role in Drp1 regulation and membrane remodeling. Moelcular dynamics simulations supported the experimental indications of an IDP structure and uncovered specific assembly sites on the protein globule surface. We are currently engaged in writing a manuscript for publication with our experimental collborators.

The final study presented in Chapter 4 is an attempt to provide more insight into the nature of how the assembly of initially disordered proteins is correlated with induced order in these proteins. In this work, I studied possible mechanisms for the formation of intermolecular beta sheet content in assemblies of initially disordered silk proteins. In silk worms and spiders, such proteins begin as unstructured and highly concentrated in the gland, but undergo

noticeable structural transitions during extrusion to form strong fibers with a high content of intermolecular beta-sheet secondary structure. Experimental work on reconstituted silkworm fibroin solutions by our collaborators, Dr. Benjamin Partlow and Prof. David Kaplan, indicated a possible role of tyrosine residues in the fibroin sequence in templating this beta-sheet structure. This stimulated my simulation studies of the role of tyrosine interactions on beta-sheet formation and stability in silk-mimetic polypeptide sequences that this reported in this chapter. Together with our experimental collaborators, we examined the role of the tyrosine residues in templating of beta-sheet structure in fibroin solutions, partial results of which were published in *Biomacromolecules*.

Also in collaboration with the Kaplan Group at Tufts University, I had an opportunity to work with Dr. Lin Yinan, who designed an innovative peptide that sharply responds to solution pH as a triggering element in a responsive biomaterial system. This peptide makes a sharp transformation from random-coil structure into an amphiphilic helical structure that forms a coiled-coil assembly upon decreasing pH to below 5. Using molecular dynamics simulations, I studied this cooperative transformation and compared possible coiled-coil oligomer states using free energy calculations. A summary of this work is offered in Appendix A), which will be included as part of a larger joint computational/experimental manuscript for submission to *ACS Biomaterials*.

Throughout my Ph.D. research, I developed various *python* or *TCL* scripts for building initial molecular configurations and for extracting and analyzing the simulation results. One of the challenges in protein simulation is to define appropriate initial molecular structures, particularly for disordered proteins where the 3D structure is not experimentally known. To resolve this issue, I developed a python package to design polypeptide chain in desirable conformation for a given sequence. A graphical user interface (GUI) for this code was also developed under my guidance by an undergraduate student, Emilie Fortin. We are currently working to implement it in the popular molecular VMD molecular visualization software package [52]. This script is presented in Appendix B.

CHAPTER 2

Mechanics of Intrinsically Disordered Proteins

2.1 Background: Lessons from Polymer Physics

In this chapter, the structure and elasticity of a de novo gluten-mimetic disordered polypeptide were simulated and investigated using models introduced in polymer physics. Proteins are polymers build up by amino acids linked into a polypeptide chain. Therefore polymer physics models can be employed to describe the global structure and mechanics of the polypeptide chain. In this first section, a short introduction is provided on single polymer models and their applications.

2.1.1 Ideal chain and entropic spring

The simplest possible model presents a polymer as an *ideal chain*, where there is no net interaction (repulsive or attractive) among the monomers which are linearly linked to each other by a constant length bond. At a given conformation, the polymer end-to-end distance vector $\overrightarrow{R_e}$, containing N monomers with the bond length of l, can be written as, $\overrightarrow{R_e} = \sum_{i=1}^{N} \overrightarrow{r_i}$, where monomer i and $i+1$ are linked by the vector $\overrightarrow{r_i}$, with $|\overrightarrow{r_i}| = l$ (See Figure 2.1 A). Although N monomers are connected by $N-1$ bonds in a linear chain, the number of bonds is traditionally approximated by N ($N-1 \approx N$, since $N \gg 1$). Thus, an average squared end-to-end distance over all possible conformations, $\langle R_e^2 \rangle$ is given by,

$$\begin{aligned} \langle R_e^2 \rangle = \langle \overrightarrow{R_e} \cdot \overrightarrow{R_e} \rangle &= \langle \left(\sum_{i=1}^{N} \overrightarrow{r_i} \right) \cdot \left(\sum_{j=1}^{N} \overrightarrow{r_j} \right) \rangle = \sum_{i=1}^{N} \sum_{j=1}^{N} \langle \overrightarrow{r_i} \cdot \overrightarrow{r_j} \rangle \\ &= Nl^2 \end{aligned} \tag{2.1}$$

where $\langle \overrightarrow{r_i} \cdot \overrightarrow{r_j} \rangle = 0$ for $i \neq j$ since there is no correlation between the directions of the bonds in the absence of an external force for an ideal chain. Indeed, the direction of each bond is completely independent of any other bonds. This requires considering monomers as point particles which can be located anywhere, even on top of each other at the same position. Therefore, an ideal chain is equivalent to a simple random walk in three dimensions. In contrast, a chain consisting of monomers with excluded volume interactions can be described by a self-avoiding random walk model.

The three-dimensional random walk model can be generalized from the results in one dimension. In one dimension, there are only two choices for each step: going in the positive

or in the negative directions. Each successive step is independent of the previous ones. Thus, the probability of arriving at a distance x from the starting point after N steps of unit size l is given by,

$$P(N, n_+) = \frac{1}{2^N} \frac{N!}{n_+! n_-!}$$

where n_+ and $n_- = N - n_+$ are the number of steps in the positive and the negative directions, respectively. Using $x = (n_+ - n_-)l$ and $N = n_+ + n_-$, we can write n_+, and n_- as functions of N and x in the expression for P. For the case of small net end-to-end distance, $x \ll Nl$, the Stirling approximation, $m! \approx (m/e)^m \sqrt{2\pi m}$ for $m \gg 1$, allows the probability expression to be approximated by a Gaussian function,

$$P(N, x) = \frac{1}{\sqrt{2\pi N l^2}} \exp\left(\frac{-x^2}{2Nl^2}\right)$$

This relation can be generalized for three dimensions to obtain $P(N, \overrightarrow{R_e})$,

$$P(N, \overrightarrow{R_e}) = \left(\frac{3}{2\pi N l^2}\right)^{3/2} \exp\left(\frac{-3R_e^2}{2Nl^2}\right) \tag{2.2}$$

By inspection, the second moment of the distribution is $\langle R_e^2 \rangle = Nl^2$, which serves as a characteristic average chain size, while the maximum end-to-end distance is $L_c = Nl$, also known as the *contour length*. Such an ideal chain is also referred to as a Gaussian chain.

In the absence of interactions between monomers, the free energy of an ideal chain is determined exclusively from its conformational entropy, which can be obtained from Boltzmann's relation, $S(N, R_e) = k_B \ln \Omega(N, R_e)$, where k_B is Boltzmann's constant and Ω_i is the number of ways that the chain can have an end-to-end distance R_e. Since $\Omega(N, R_e) \propto P(N, R_e)$, using Equation 2.2, the entropy and free energy of an ideal chain can be expressed as a function of the chain end-to-end distance, R_e, as

$$\begin{aligned} S(N, R_e) &= -k_B \ln \left(\frac{3}{2\pi N l^2}\right)^{3/2} \exp\left(\frac{-3R_e^2}{2Nl^2}\right) + S_0 \\ &= k_B \left(\frac{-3R_e^2}{2Nl^2}\right) + S_0' \\ F(N, R_e) &= -TS = \frac{3}{2} k_B T \left(\frac{R_e^2}{Nl^2}\right) + F_0' \end{aligned} \tag{2.3}$$

where T is the temperature. Furthermore, the force required separate the ends of ideal chain can be obtained from the free energy expression in Equation 2.3 by using Equation 2.1 to obtain

$$f = \frac{\partial}{\partial R_e} F(N, R_e) = 3k_B T \frac{R_e}{Nl^2} = \frac{3k_B T}{Nl^2} R_e \tag{2.4}$$

which illustrates that an ideal chain acts as an entropic spring with spring constant, $k = 3k_B T/Nl^2$. Note that the above linear relation between force and end-to-end distance is only valid for small extension ($R_e \ll Nl$), where the chain still behaves approximately as a Gaussian chain. For large extension, the principal assumption in the ideal chain model, having uncorrelated bonds, is not valid. The force-extension relation of an ideal flexible chain for arbitrary extension can be calculated in the freely-jointed chain (FJC) model. Alternatively, for chains that are not intrinsically flexible, the worm-like chain model (WLC) can be used to accurately describe force-extension relation for such so-called semi-flexible polymers.

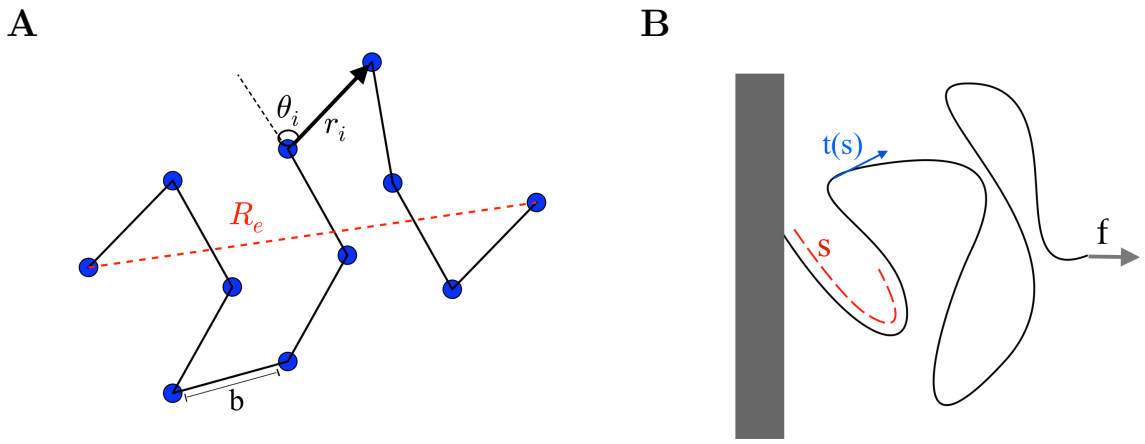


Figure 2.1: (A) Ideal chain model where chain segment i was linked to the next segment via vector r_i with length l and angle θ_i. R_e indicates end-to-end distance. In the continuum limit, where l is much smaller than other characteristic length scales, the angular increments between steps become small enough to describe the chain as a smooth curve in a so-called worm-like chain model. (B) Schematic diagram of a worm-like chain under tension f, indicating a tangential unit vector $\mathbf{t}$ at a curvilinear distance s along the chain.

2.1.2 Polymer elasticity models: worm-like chain

The elasticity of a polymer for a large extension follows a nonlinear relationship which has been commonly described by two models of polymer elasticity, known as the freely-jointed chain (FJC) and the worm-like chain (WLC) models. The FJC model is derived by treating the ideal chain as a series of bonds that point in any direction with equal probability, independently of neighbouring bonds. By considering the correlation of bonds between monomers due to the application of external force at finite temperature, an exact expression for the force-extension relationship is obtained:

$$R_e = Nl\mathcal{L}(fl/k_BT) \tag{2.5}$$

where $\mathcal{L}(x) = \coth(x) - 1/x$ is the Langevin function. This model only accounts for the entropic contribution to the restoring force due to a reduction in the number of available configurations under tension and therefore is only strictly valid for ideal chains. In addition to entropic contributions, the WLC model also implicitly includes enthalpic effects by introducing a bending energy to the chain to account for the semi-flexible nature of realistic polymers.

The WLC model can be described using a version of the ideal chain model known as the freely rotating chain (FRC). In the FRC model the polymer is described by a continuous chain of N bond vectors, with length l, which each of which is at an angle θ to the adjacent segment. Thus, the average projection of adjacent segments, $(1/l)\langle \overrightarrow{l_i} \cdot \overrightarrow{l_{i+1}} \rangle$ is equal to $l\cos(\theta)$ (whereas it is zero for an ideal chain). Likewise, the projection of a distant segment $i + j$ onto the segment i direction is $l\cos^j(\theta)$. Therefore, the length scale beyond which segment orientational correlations are negligible, can be obtained by projection over the next n segments ($n \gg 1$),

$$p = l\sum_{i=0}^{n}\cos^j(\theta) = l\frac{1-\cos^n(\theta)}{1-\cos(\theta)} \approx \frac{l}{1-\cos(\theta)} \tag{2.6}$$

where p, the so-called *persistence length*, is a measure of polymer stiffness. Flexible chains have small persistence length because the correlations between segments are lost quickly. In contrast, the persistence length of rigid rod-like chains is the entire contour length of the chain Nl. The most realistic case is the so-called semi-flexible chain, where $l \ll p \ll Nl$, and

the chain is stiff on intermediate length scales.

Using the FRC in the limit of small θ provides an approximation to the WLC behaviour. The generalization of Equation 2.1 for the end-to-end length of flexible ideal chains to the case of the FRC model is obtained from

$$\langle R_e^2 \rangle = \sum_{i=1}^{N} \sum_{j=1}^{N} \langle \overrightarrow{l_i} \cdot \overrightarrow{l_j} \rangle = l^2 \sum_{i=1}^{N} \sum_{j=1}^{N} \cos^{|i-j|}(\theta)$$

In the continuum limit of a highly directed FRC, θ is small and therefore $\cos(\theta)$ can be approximated by $(1 - \theta^2/2)$ in Equation 2.6, giving $l/p = \theta^2/2$. By replacing $\cos(\theta)^{|i-j|} \approx \exp(-|i-j|\theta^2/2) \approx \exp(-l|i-j|/p)$ in the above equation and changing the sum over segments into an integral over contour length ($l \sum_{i=1}^{N} \mapsto \int_0^{L_c} ds$) the end-to-end distance becomes,

$$\begin{aligned} \langle R_e^2 \rangle &= l^2 \sum_{i=1}^{N} \sum_{j=1}^{N} \exp\left(\frac{-|i-j|l}{p}\right) \\ &= \int_0^{L_c} \left[\int_0^{L_c} \exp\left(\frac{-|s-s'|}{p}\right) ds \right] ds' \\ &= 2pL_c - 2p^2 \left(1 - e^{-L_c/p}\right) \end{aligned} \tag{2.7}$$

Note that for the $p \ll L_C$ limit, $\langle R_e^2 \rangle \simeq 2pL_c$ which is identical to the ideal chain result; consequently from Equation 2.1, $p = l/2$.

There is an energy associated with the bending in the chain conformation. In absence of an external force, the bending energy for a WLC with a bending elasticity, $\kappa \sim pk_BT$, is given by

$$E_{bend} = \frac{\kappa}{2} \int_0^{L_c} \left| \frac{d\mathbf{t}}{ds} \right|^2 ds$$

where s is the distance along the curve and $\mathbf{t}(s)$ is the tangent vector along the chain (See Figure 2.1 B). Stretching the chain then does work against bending energy. The Hamiltonian for a WLC under a tension f in the direction of extension z can be written as

$$H = \int_0^{L_c} \frac{p}{2k_BT} \left| \frac{d\mathbf{t}}{ds} \right|^2 ds - fz$$

Using the Boltzmann distribution, $P(f,z) \propto exp(-H(f,z))$ with the Hamiltonian expression above, the average end-to-end distance as a function of external force can in principle be calculated via the partition function $\mathcal{Z}(f,z)$. Marko and Siggia [53] proposed a simple expression for force-extension behavior of a WLC that is a simple interpolation between the high and low force regimes, each of which can each be calculated analytically. The resulting approximate force-extension relation for the WLC is given by,

$$\frac{fp}{k_B T} = \frac{1}{4}\left[\left(1 - \frac{z}{L_c}\right)^{-2} - 1\right] + \frac{z}{L_c} \tag{2.8}$$

which has been shown able to describe fairly accurately the force-extension behaviour of semi-flexible biomolecules such as DNA.

2.1.3 Real polymer size in solution and scaling-law

Simple polymer physics models are able to predict the dependence of polymer size on the number of monomer units without requiring detailed knowledge about long range interactions between monomers or interactions between solute. Polymer sizes are usually reported in terms the time averaged end-to-end distance R_e or the time-averaged of radius of gyration (R_g), which is given by

$$\langle R_g^2 \rangle = \langle \frac{1}{N} \sum_{i=1}^{N} (\mathbf{r}_i - \mathbf{r}_{cm})^2 \rangle$$

where $\mathbf{r}_i$ indicates position of monomer i and $\mathbf{r}_{cm}$ is according to the center of mass position. The average radius of gyration is proportional to the average end-to-end distance of the chain. For an ideal flexible polymer, one finds $\langle R_g^2 \rangle = (1/6)Nl^2$. Furthermore, this power-law relationship between the radius of gyration, R_g, of the polymer and its number of monomer (polymer subunits), N, can be extended to more realistic polymer behaviour observed experimentally in solution via,

$$R_g \propto N^{\nu} \tag{2.9}$$

where the value for the power exponent, ν, is sensitive to both the intra-chain monomer and monomer-solvent interactions.

Although a real polymer does not behave fully as an ideal chain in most conditions, there is a length scale below which the polymer chain may be considered to be unaffected by other factors, whether they are external forces, excluded volume effects, or solution interactions. Below such a length scale, the thermal contribution to the free energy is dominant and the polymer exhibits ideal behaviour. Therefore, the polymer chain can be divided into correlation blobs of size ξ, within which ideal behaviour is obtained. For such a correlation blob with g monomers per blob, the size ξ of the blob obeys ideal chain statistics: $\xi \approx bg^{1/2}$. Overall, the chain can be thought of as grouping of $n = N/g$ such thermal blobs. Above the length scale ξ, the real interactions between monomers and solvent molecules are dominant and we obtain the scaling behavior appropriate for the actual solvent conditions (e.g. good and poor). Thus a super-chain formed of blobs of size ξ has a characteristic global size given by $R_g \simeq n^{\nu}\xi \sim N^{\nu}$. In the following, the behaviour observed for various solvent conditions is summarized.

In *poor solvent*, attractive forces between monomers are dominant and monomer-solvent interactions are not preferred. Thus, polymer self-interaction overrides entropic considerations, resulting in blobs that are closely packed into a compact spherical shape to minimize the surface contact with the solvent with the polymer. In this case, the scaling law for the radius of gyration can be calculated as,

$$\frac{4}{3}\pi R_g^3 \approx n\frac{4}{3}\pi\xi^3, \mapsto R_g^3 \approx (N/g)(b.g^{1/2})^3$$
$$R_g \propto N^{1/3}$$

with the scaling exponent of $\nu = 1/3$.

In absence of preferential interaction for the chain monomers, where monomer-monomer interactions are as favored as monomer-solvent interactions (so-called *θ-solvent* conditions), the chains behave as ideal random walks on all length-scales and the ideal chain statistics are recovered. Thus, the scaling exponent can be obtained from the power-law relation between the ideal chain and its number of monomer (blobs),

$$R_g^2 \approx n\xi^2, \mapsto R_g^2 \approx (N/g)(b.g^{1/2})^2$$
$$R_g \propto N^{1/2}$$

In *good solvent*, the monomers preferably interact with solvent molecules rather than

other monomers, thus overlapped blobs are avoided and the superchain of blobs behaves as self-avoiding random walk. In this state, the swelling effect due to the excluded volume for the blobs is counterbalanced by the entropy loss of the extended chain conformations. Using this analogy, Flory [54] developed a simple free energy model that predicts a power-law relation between polymer size and number of monomers with a higher exponent, $\nu = 3/5$,

$$R_g \propto N^{3/5}$$

These simple polymer models for equilibrium conformation and response to applied force, were originally developed for simple homopolymers. However, some insight has been gained by applying these ideas to more complex biopolymers, such as proteins. For instance, the scaling behaviour of the radius of gyration in proteins provides indirect information about protein solubility and has been used to distinguish folded proteins from denatured/disordered ones. Also, using polymer models for chains under tension, force-extension experiments on proteins and DNA have provided valuable insight to the local stiffness of these biopolymers. To motivate the simulation studies presented in this chapter, the possible conformational and mechanical properties of disordered proteins are next discussed in terms of such simple polymer models.

2.2 Introduction

Lack of well-defined three-dimensional structure enables disordered proteins/regions to fluctuate between a large number of conformational ensembles and in some case to behave as *entropic springs*. This concept from polymer physics, described in the previous section, is due to the reduction in conformational entropy that occurs during stretching or compressing of a disordered region. In such a case, a disordered chain will attempt to respond by re-populating higher entropy states and thus will provide a restoring force to do so. Elastomeric proteins are examples of disordered proteins that exhibit this classic entropic spring behaviour.

Elastomeric proteins are present in a wide range of biological systems, where they are employed for their toughness and flexibility. These proteins are able to undergo large deformation without rupture, partially store the energy involved in deformation, and then recover their original state when stress is removed. Titin and elastin are two well-known examples among widely distributed elastomeric proteins in the animal kingdom. The elastic properties of titin contribute considerably to the passive restoring force in muscle, and elastin is a basic component of soft tissues such as skin and blood vessels providing material toughness. Although elastomeric proteins consist of diverse amino acid sequences, they normally have a substantial unstructured component, often comprised of repetitive sequences, which are responsible for their remarkable elasticity [55]. Several examples of elastomeric proteins along with their elastic domain repeat motifs are given in Table 2.1.

Protein	Elastic repeat motifs
Titin	PPAKVPEVPKKPVPEEKVPVPVPKKPEA
Elastin	VPGG / VPGVG/ APGVGV
Resilin	GGRPSDSYGAPGGGN / GYSGGRPGGQDLG
Glutenin HMW subunit	PGQGQQ / GYYPTSPQQ

Table 2.1: Repeat motifs for elastomeric proteins [55, 56].

Several studies have probed the mechanical properties of biological molecules such as dsDNA [57] and proteins [58], including some of the elastomeric proteins [59], using single molecule experiments and molecular simulations. Traditionally, scientists utilized the worm-like chain (WLC) model [53] to analyze the force-extension data from their studies.

The worm-like chain (WLC) model was originally used in polymer physics to describe the behaviour of semi-flexible homopolymer polymers. In this model, reviewed in the previous

section, the extension of a polymer under the force, f, is related to the total length, contour length L_c, and persistence length (p) of the polymer in the approximate relation [53],

$$\frac{fp}{k_BT} = \frac{1}{4}\left[\left(1 - \frac{z}{L_c}\right)^{-2} - 1\right] + \frac{z}{L_c} \tag{2.10}$$

where z is the end-to-end distance, k_B is the Boltzmann constant, and T is absolute temperature in kelvin ($k_BT \sim 4.3$ pN.nm at 310 K). The WLC is a monolithic model in which the force-extension profile for a polymer is described by a single persistence length along the entire polymer, whereas biopolymers often have a heterogeneous composition and therefore non-uniform mechanical properties. Therefore, the WLC model often fails to explain the elasticity of the proteins within which the local stiffness of the protein varies significantly. Moreover, this model is unsuccessful at large forces where chemical bonds of the polymer are over-stretched, modifying the intrinsic elasticity of the chain [60].

Distinct elastic properties, as well as a repetitive sequence (low sequence complexity), make elastomeric proteins a great candidates for various applications in protein engineering and biomaterials science. Despite numerous studies for many years, the precise nature of their elastic properties has been a point of contention, with competing theories proposing elastic properties as results of loss of entropy [61], hydrophobic interactions [62], and mechanics of transient turn structures [63, 64].

This chapter presents simulation studies the origin of distinct elasticity for wheat glutin mimetic (WGM) polypeptides inspired by the repeat motifs in the native Glutenin protein. A series of MD simulations of a WGM polypeptides in an explicit aqueous solvent system were employed to reveal some key molecular features of the elastic properties of this disordered polypeptide. First, the equilibrated conformation of this designed polypeptide was examined starting from fully extend conformations. Then, the force-extension profile for the polypeptide was obtained for steered MD simulations and compared with predictions of the WLC model. The observed complexity in local rigidity was found not to be consistent with the WLC model predictions. Moreover, local properties that may cause this complicated elastic behaviour were reviewed for different force/extension regimes, where predominant factors responsible for the elastic response such as dihedral angles, hydrophobic interactions, and irregular hydrogen bonds were explicitly identified and discussed.

2.3 Simulation and Analysis Methods

2.3.1 Wheat-gluten mimetic polypeptide sequence

Glutenin is a member of the wheat-gluten protein family which are responsible for viscoelastic properties of a dough [65]. Glutenin consists of high molecular weight (HMW) and low molecular weight (LMW) subunits. In particular, central repetitive sequences of HMW confer elasticity to the protein [66]. The HMW contains short repeat motifs (PGQGQQ and GYYPTSPQQ), which account for more than 90% of this domain [67]. Based on these short HMW repeat motifs, a wheat-gluten mimetic (WGM) peptide was designed in the Harden group at the University of Ottawa, where it was cross-linked to formed a polypeptide hydrogel biomaterial with high elasticity. This WGM polypeptide contains eleven identical motifs,

P G Q G Q P E Y D P T S L Q E

In the above motif sequence, negatively charged glutamic acid (E) and aspartic acid (D) residues along with polar residue such as glutamine (Q) enhance the solubility of the WGM polypeptide. On the other hand, proline and glycine amino acid influence rigidity and flexibility of the polypeptide, respectively. These residues also favour disordered conformations in solution. A tyrosine (Y) residue was also included in the motif to allow for the polypeptide to be chemically linked together and form a macroscopic hydrogel network (via the formation of intramolecular di-tyrosines). The full sequence of the WGM polypeptide, including nonrepetitive N- and C-terminal regions, contains 216 amino acids as follows,

MA-PGQGQCEYDPTSLQE-$[$PGQGQPEYDPTSLQE$]_{11}$-PGQGQPEYDPTSLQC-
PGQGQPKYAAALEHHHHHH

The conformation and elastic properties of this WGM sequence were investigated by series of simulations in this work.

2.3.2 Equilibration to a compact structure

An extended structure of the WGM polypeptide was generated by setting dihedral angles, (ϕ, ψ) to (-120, 135) for all residues using a custom designed Python code developed by the author (See Appendix B). Starting from this extended structure, with an initial 726 Å end-to-end distance and no regular secondary structure, the WGM polypeptide was equilibrated

in series of simulations, in which the final configuration of one simulation was used as an inital conformation to start the next one in a relatively smaller simulation box (since the polypeptide conformation progressively shrank during the simulations). The polypeptide reached a steady state compact but disordered conformation well within a total simulation time of $\sim 1\ \mu s$. In particular, after 750 ns of simulation time the radius of gyration reached a stable value. All statistical analysis of data were performed during the last 50 nanoseconds of the simulation (from 950 ns to 1000 ns).

2.3.3 Force-extension measurements

In the previous simulation sets, the WGM polypeptide shrank from the extended structure to a configuration with an end-to-end distance of ~ 360 Å after about 10 ns. This configuration was extracted and used to study the WGM polypeptide behavior under tension in sets of steered MD simulations, where the C_α of the first residue (MET1) was fixed and a constant force was applied to the C_α of the last residue (HIS216) along the z-direction. A series of forces, ranging from 7 pN to 209 pN were applied to this initial conformation and two simulations were carried out for each force. Simulations were executed for 70 ns to 90 ns, a time sufficient for plateauing of the end-to-end distance to be observed (for 209 pN, steady state conditions were achieved after 70 ns but for 35 pN only ~ 10 ns was required). Values for the end-to-end distance converged in both simulations for all forces, except for the very low force simulations at 7 pN.

2.3.4 General simulation details

In all simulations, peptide chains were solvated within a neutral water-ion solution (0.2 M NaCl) with periodic boundary conditions using the water box module of VMD [52]. The simulation box was extended to maintain a minimum of 15 Å between the peptide atoms and the boundary of the simulation box. All MD simulations employed a Langevin thermostat and barostat and were conducted under the isothermal-isobaric ensemble (NPT) at 300 K and 1 bar. The CHARMM22 force field [68] with the protein backbone improvement [69] was adopted for the peptides. The TIP3P model was used for the water molecules in which partial charges were assigned to three interaction sites, the positions of water hydrogens and oxygen atoms, to mimic the polarity of water molecules [70]. The electrostatic interactions were computed via the particle-mesh Ewald (PME) algorithm in which the interaction potential

is given as sum of two series: a short-range and a long-range series. The short-range series is applicable for small distances and converges in real space. On the other hand, the long-range part converges in Fourier space, where the charge density is evaluated on a mesh. A cutoff of 12 Å was set for Lennard-Jones interactions and the real-space part of PME. The time step was set at 2 fs and all simulations were carried out with the NAMD 2.9 software package [71].

2.3.5 Structural analysis

Using the polypeptide trajectories from the last nine nanoseconds of the simulations, quantitative analysis of the WGM structure was achieved by comparing the secondary structure content, dihedral angle distributions, and intramolecular interactions within the repetitive central sequence. These examinations were performed for elongated WGM polypeptide at different forces, as well as for compact conformations obtained at zero external force.

All secondary structure content was measured using the STRIDE algorithm [72]. This algorithm assigns secondary structures based on combined parameters, such as hydrogen bond energy, statistically driven backbone dihedral angle criteria, and hydrogen bond patterns. Helical structures, such as the 3_{10}-helix, α-helix, and π-helix, are recognized by having at least two consecutive i-i+3, i-i+4, and i-i+5 hydrogen bonds, respectively. A minimal beta-sheet is identified by two consecutive inter-strand hydrogen bonds, while isolated beta-bridges were defined by single inter-strand hydrogen bond pair.

Dihedral angle data, ϕ and ψ, spanning a range from -180° to 180°, was divided into 250 bins, and dihedral angle distributions were obtained by averaging number of angles that fall into each bin over last 9 ns of the simulations. The results were then normalized to find the probability of having an angle within individual bins. A smoothed probability distribution was obtained by translating a window of 10 bins width continuously over the results, averaging the observations that fall in the window.

Intramolecular contact in the structures were quantified in terms of hydrogen bonds and hydrophobic interaction. A hydrophobic association event was noted when the distance between the sidechain centers of mass for two hydrophobic residues (proline and leucine in WGM) was less than 6 Å. A hydrogen bond was defined by criteria of special separation and orientation of an atom with a covalent hydrogen bond (the donor, D) and another electronegative atom (the acceptor, A). A hydrogen bond, D-H...A, was noted when the distance between D and A was less than a cut-off distance (set to 3.2 Å) and when the angle,

defined as the deviation of Donor-Hydrogen-Acceptor from linearity, was less than the cut-off angle (set to 40°). Note that in an ideal hydrogen bond, the donor and acceptor are co-linear with a bond angle of 0° based on the angle definition used here.

The radius of gyration was calculated using coordinates of C_α atoms in polypeptide backbone. Comparisons were made with natively ordered proteins obtained from Protein Data Bank (PDB bank) database [20] with lengths between 20 and 600 residues and with less than %50 sequence identity as defined in the PDB bank.

2.4 Results and Discussions

2.4.1 Structural aspects of the compact state

Starting from an extended configuration, the wheat-gluten mimetic (WGM) polypeptide was equilibrated in explicit water-ion solution over 1000 ns, during which it reached a steady state conformation. The polypeptide radius of gyration, R_g, as a function of time, along with two snapshots indicating its configuration in the first and last few nanoseconds of the simulation, are shown in Figure 2.2. The steady-state, compact conformation was achieved after about 750 ns, wherein its R_g reached a plateau with an average value of 23 ± 1 Å (averaged over last 250 ns). The spatial configuration of this compact structure was further characterized by the *principal moments* of the radius of gyration (the elements of diagonalized gyration tensor), which were found to be $\sim$ 18 Å, 11 Å, and 8 Å. Therefore, the compact structure adopted a somewhat irregular and open ellipsoidal shape, in contrast to natively folded proteins that are often found in more compact globular structures. Interestingly, the R_g value for WGM polypeptide was larger than typical values for natively folded proteins of the same number of residues and also smaller than values for chemically denatured proteins of the same number of residues. In Figure 2.2 B, R_g for the WGM polypeptide (indicated as a point) is presented along with R_g data for sets of natively folded and chemically denatured proteins depicted as a function of the number of residues, showing that the behaviour of the WGM polypeptide is intermediate to that of folded and denatured proteins.

Generalized Flory theory [54] predicts that radius of gyration of a polymer, R_g, scales with its number of monomers (polymer subunits), N, via $R_g \propto N^\nu$, where ν is scaling component depending on interactions between polymer chain segments and solvent molecules. This relation was fit to the data in Figure 2.2 B, where the corresponding the scaling component

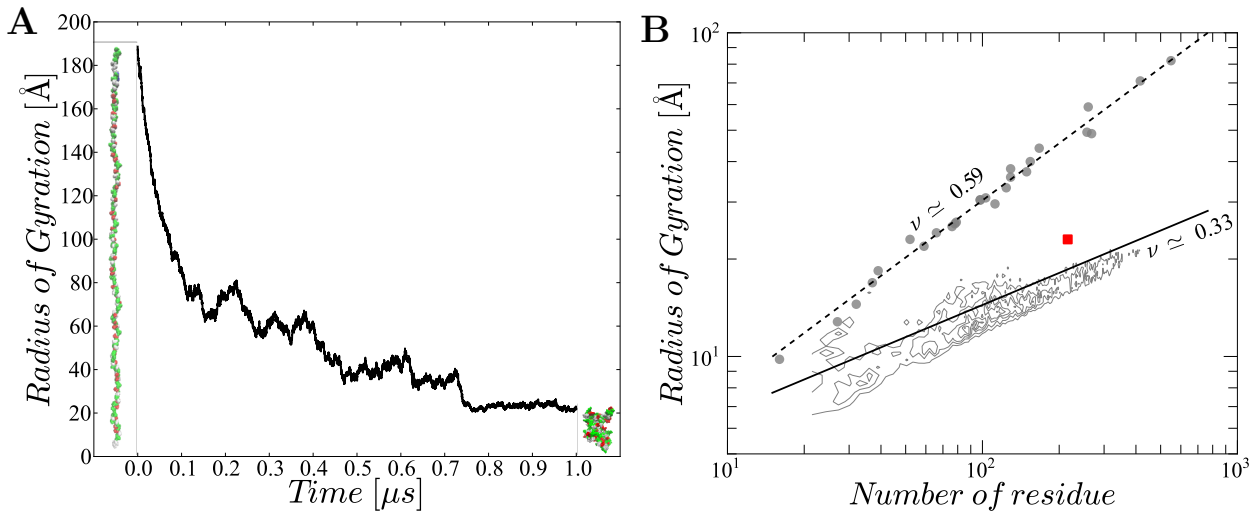


Figure 2.2: (A) Radius of gyration (R_g) of wheat-gluten mimetic polypeptide was measured as a function of time using backbone C_α coordinates during 1000 ns simulation where polypeptide shrank from extended conformation (snapshot on the left) to its compact state (snapshot on the right). (B) R_g as a function of the number of residues (N) for chemically unfolded proteins (gray circles), the WGM polypeptide (red square), and natively folded proteins (distribution represented in contour plot in gray line). R_g values for unfolded and folded proteins were taken from Kohn et al [73] and from 12219 native proteins downloaded from Protein Data Bank [20]; respectively. Black solid and dashed lines are fits according to $R_g \propto N^\nu$ and resulting scaling exponents are indicated.

value is indicated. The ν value for natively ordered protein is $\sim 1/3$, the theoretical value for spherical compact globule, while for chemically denatured protein it is close to $\sim 3/5$, the theoretical value for a self-avoiding random walk. Therefore, the WGM polypeptide is not as compact as natively folded proteins and is not as swollen as denatured proteins. Moreover, a more careful analysis indicates that the WGM polypeptide conformation is more compact that an ideal chain, possibly due to favorable intramolecular interactions between hydrophobic residues. As such, it has characteristics of globular Intrinsically disordered proteins (IDPs), which do not fold to a regular compact structure yet do not completely behave as a random coil. Although IDPs are found in the same physiological conditions as natively folded proteins, their fairly distinct scaling exponents suggest different residual interactions for these proteins, due to particular characteristics of the protein sequence (such as hydrophobicity, charge, size of side-chain, and sequential position of amino acids). In Figure 2.3 A, a final compact conformation of WGM is illustrated at the residue level where amino acids are presented in

different colours according to their physiochemical characteristics.

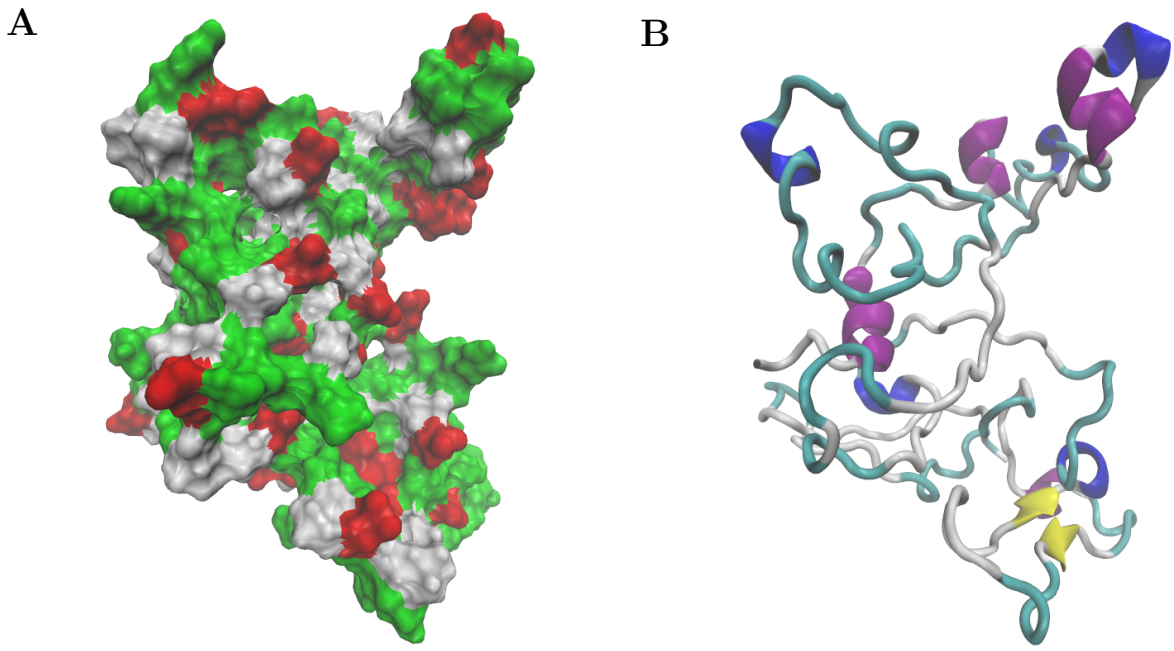


Figure 2.3: Snapshot of the compact state of the wheat-gluten mimetic polypeptide. (A) Schematic of the polypeptide at residue level (green, polar residues; white, hydrophobic residues; red, negatively charged residues) at the end of 1000 ns of simulation. Solvent molecules are not represented. (B) Schematic of secondary structures of the polypeptide backbone (white tube, coil; cyan tube, turns; blue ribbon, 3_{10}-helix; purple ribbon, α-helix; yellow arrows, β-sheet). Secondary structures were determined using STRIDE methods [72].

The disordered characteristics of WGM are more apparent in the secondary structure content of the polypeptide. Ordered proteins have a high content of regular secondary structure, such as alpha helix and beta sheet, indicative of coordinated backbone hydrogen bonds, which are packed together to form fully compact globular structures. However, the amount of secondary structure in the compact state of the WGM polypeptide is significantly low (See Table 2.2) compared to folded proteins. Indeed, more than 87% of the WGM polypeptide consists of irregular secondary structures (turns and coil) which do not have regular hydrogen bonds and which are more flexible than regular secondary structures. This is summarized in the snapshot of the WGM backbone in Figure 2.3 B, where secondary structures of the WGM polypeptide are indicated by different colours. In the next two sections, the elastic properties of WGM and its relation with these local structures were investigated.

	Helical (%)	Beta (%)	Turn (%)	Coil (%)
compact state	9.0 ± 1.1	3.0 ± 0.3	43.5 ± 1.3	44.5 ± 1.5

Table 2.2: Secondary structure content of the compact state of the wheat-gluten mimetic polypeptide. Secondary structure contents of the molecule were measured in four categories: helical (α-, π-, and 3_{10}-helix), Beta-sheet, Turn, and Coil conformation which all were assigned using STRIDE methods [72] (See Section 2.3.5 for more details). Statistical analysis was made using simulation data in the last 50 ns, corresponding to steady-state conditions, that has been divided into 5 batches of 10 ns each to calculate averages and standard deviations.

2.4.2 Stretching the WGM polypeptide

The WGM polypeptide amino acid sequence was inspired by an elastic domain in Glutenin, a member of elastomeric protein family. While these proteins are known for their extraordinary flexibility, there is less known about underlying molecular mechanisms responsible for their elasticity. Herein a series of steered molecular dynamic simulations were conducted to explore the elastic response of WGM to applied tension and its relationship to the structural disorder of this polypeptide.

A configuration of WGM with end-to-end distance of 360 Å, selected from the previous simulation data, was equilibrated under a wide range of tensions (low tensions: 7 pN, 20 pN, 35 pN; intermediate tensions: 97 pN, 111 pN; and high tensions: 174 pN, 209 pN) in series of simulations. Two separate simulations were carried out for each tension to check for reproducible response. The end-to-end distance of the polypeptide as a function of tension was reproducible in all simulations except for those conducted at a 7 pN tension. For this low tension, not only did the end-to-end values for two simulations differ, but they also failed to reach steady state values during the simulations, possibly due to issues related to the metastability of numerous intermediate conformations sampled during extensional evolution. For all higher tensions, steady-state values of the end-to-end distance were obtained in the simulations and time-averaged end-to-end distance values were measured over 15 ns in steady-state conditions to obtain a force-extension profile of the polypeptide. Examples of the molecule trajectories under tension are given in Figure 2.4 A.

The elastic response of the WGM polypeptide for the range of external forces was represented in terms of the force-extension profile in Figure 2.4 B, where its elasticity response was analyzed in terms of the worm-like chain (WLC) model. The standard worm-like chain

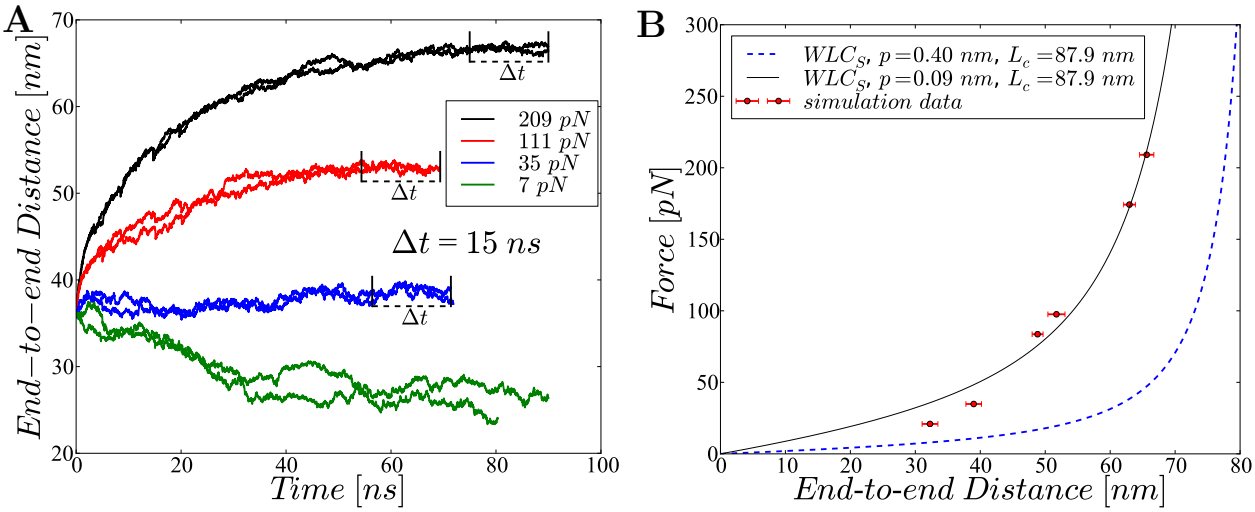


Figure 2.4: Force spectroscopy of wheat-gluten mimetic polypeptide. (A) End-to-end distance as a function of time for WGM polypeptide under various tension. (B) Force-extension data point extracted from simulations (red circles, average; red bars, variation), alongside fitted WLC model (black curve). For comparison, a typical force-extension profile for an ordered protein, as predicted by the WLC model with a persistence length of the peptide size is shown by the dashed blue curve.

(WLC_S) model, using Equation 2.10, was fit to the average end-to-end at difference forces using $L_c = 87.9$ nm and $p = 0.09$ nm as contour length and persistence length parameters. The fit contour length value is in very good agreement with 86 $\pm$ 4 nm, the expected value for the WGM polypeptide (using 4.0 $\pm$ 0.2 Åas the characteristic length of a single peptide unit [74]). The persistence length of $p = 0.09$ nm found from the fit of the WGM data to the WLC model is also in fair agreement with the 0.18 nm value reported in force-extension experiments on elastin-like polypeptides using single molecule force spectroscopy [75].

On the other hand, computational studies have reported a persistence length of 0.4 nm [58] for forced unfolding of Ubiquitin protein (known as typical natively ordered protein). The much smaller apparent persistence length inferred from WGM simulation data is consistent with expectations for elastic disordered polypeptides, which typically have persistence lengths that are smaller than those of typical globular proteins. Thus, for any given applied tension the WGM polypeptide should be at a smaller extension than a natively ordered protein with the same number of residues. This is highlighted in Figure 2.4 B, where a hypothetical force-extension profile for a protein with the same contour length as WGM, is illustrated using the

WLC model with $p = 0.4$ nm (dashed blue curve).

Although the WLC model was able to recapitulate the overall force-extension profile for the WGM polypeptide, it failed to predict a reasonable persistence length for this polypeptide. Protein chains consist of peptide units with a size of ~ 0.4 nm; therefore, any persistence length smaller than this characteristic polypeptide segment size is not physically meaningful. There are two assumptions in the WLC model which may not be valid for these types of polypeptides. First, the WLC model assumes that the local stiffness is uniform along the chain and therefore ignores the impact of mechanical heterogeneity in the polypeptide sequences. Second, the WLC model considers that local stiffness is an intrinsic property of the system that is independent of applied forces and extensions. These are crude assumptions that may disregard essential details associated with disordered proteins such as intermolecular interactions and protein sequence heterogeneity.

The WLC model was able to approximate the elastic behavior of Ubiquitin protein with a physical value of the persistence length (0.4 nm) for those extensions of this protein that lacked secondary structure or native contact (corresponding to tensions ranging from 250 pN to 35 pN [58]). Indeed, all native contact discharged upon unfolding of this natively ordered protein. The success of the WLC model for natively ordered proteins, such as Ubiquitin, is perhaps due this lack of native contact upon unfolding. In contrast, a variety of intramolecular interactions are operative in during tension-induced WGM polypeptide extension, particularly at low tensions. As a result, the WLC model was not able to predict elastic properties of WGM accurately at low tensions. At higher tensions, sequence heterogeneity may play a role. For instance, even in absence of native contacts at high extension, protein segments having proline residues should be stiffer than other segments due to the fixed torsion angle for proline. The modified WLC model, which accounts for the elongation of polypeptide segments in situations where chemical bonds are slightly extended under strong tension, was also used to analyze the WGM force-extension profile (see Appendix 2.6.1). However, even this modified WLC model was unable to predict a consistent, physical value of the persistence length, suggesting that the mechanics of the WGM polypeptide may have distinct mechanisms of response, dictated by local structure, that are operative in different extension/force regimes. The next section considers such local mechanisms of defromation.

2.4.3 Structural aspects impacting WGM elasticity

The transient intra-molecular association and local conformation were investigated in WGM at the different applied tensions, in an attempt to explain distinct regimes of resilience/elasticity observed in simulations of this polypeptides. Local secondary structure and details of local residue interactions were first assessed for stretched and compact states, as shown in Table 2.3. For local residue interactions, the focus was on hydrophobic contacts and hydrogen bond formation. Electrostatic interactions were not considered due to the lack of oppositely charged residue in the WGM amino acid sequence. Local conformations of the WGM chain backbone were further characterized by exploring the dihedral angle distribution for individual tensions, with results illustrated in Figure 2.5. Together, Table 2.3 and Figure 2.5 reveal essential structural features of the polypeptide under tension that are used in the following discussion to propose mechanical mechanisms employed by the polypeptide at different applied tensions.

	Helical (%)	Beta (%)	Turn (%)	Coil (%)	Hbond_{BB}	HP contact
209 pN	0.0 ± 0.3	0.2 ± 0.1	1.3 ± 0.7	98.5 ± 0.8	0 ± 0	0 ± 0
111 pN	1.7 ± 0.3	0.1 ± 0.1	18.6 ± 1.6	79.7 ± 1.4	2 ± 1	3 ± 0
35 pN	3.0 ± 0.6	0.1 ± 0.1	42.1 ± 2.8	54.7 ± 2.7	5 ± 1	6 ± 1
compact	10.3 ± 0.3	2.6 ± 0.1	45.8 ± 2.2	41.4 ± 2.4	15 ± 1	18 ± 0

Table 2.3: Local structural details of wheat-gluten mimetic polypeptide in the compact and stretched states. The average secondary structural content of the WGM were measured in four categories: helical (α-, π-, and 3_{10}-helix), beta (beta-sheet and isolated bridge), turn, and coil conformations. The last two columns include numbers of backbone hydrogen bonds (Hbond_{BB}) and Hydrophobic (HP) contacts (See Section 2.3.5 for more details). Statistical analysis was made using simulation data in the last 9 ns, corresponding to steady-state conditions, that has been divided into 3 batches of 3 ns each to calculate averages and standard deviations.

At 209 pN, the molecule is completely in a coil conformation without local contacts or regular secondary structure. In the absence of sidechain association and backbone hydrogen bonding, the chain elasticity exclusively arises from the influence of force on the polypeptide chain backbone conformation. Backbone bond lengths ($N-C_\alpha$, $C_\alpha-C$, $C-N$) and bond angles ($\widehat{NC_\alpha C}$, $\widehat{C_\alpha CN}$, $\widehat{CNC_\alpha}$) are stretched by $\lesssim 1\%$ at 200 pN (according to associated potentials in the CHARMM forcefield [68]), which is significantly smaller than the 76% extension observed from 35 pN to 209 pN for the WGM polypeptide. Thus, their contribu-

tion to the chain end-to-end extension is negligible. In contrast, the ϕ dihedral angle (the $C^{i-1}N^iC^i_\alpha C^i$ torsion angle) and the ψ dihedral angle (the $N^iC^i_\alpha C^iN^{i+1}$ torsion angle) varied significantly in the high tension regime (See Figure 2.5). As tension increased, dihedral angles were shifted towards $\phi = -180°$ and $\psi = 180°$, values corresponding to fully extended backbone angles.

As a result, dihedral angles were the main contributors to the chain extension at this high tension, and their loss of freedom was responsible for elastic response of the polypeptide at high tension. In this regime, the WGM polypeptide, lacking native contacts, shares the same source of elasticity as unfolded ordered proteins, such as ubiquitin [58]. However, the WGM polypeptide has a shorter end-to-end distance due to its high content of proline. While the values of ϕ dihedral angle of other amino acids were shifted toward -180° by increasing tension, proline maintained its values centered around -60° due to its intrinsic molecular structure (See Figure 2.8, Appendix 2.6.2).

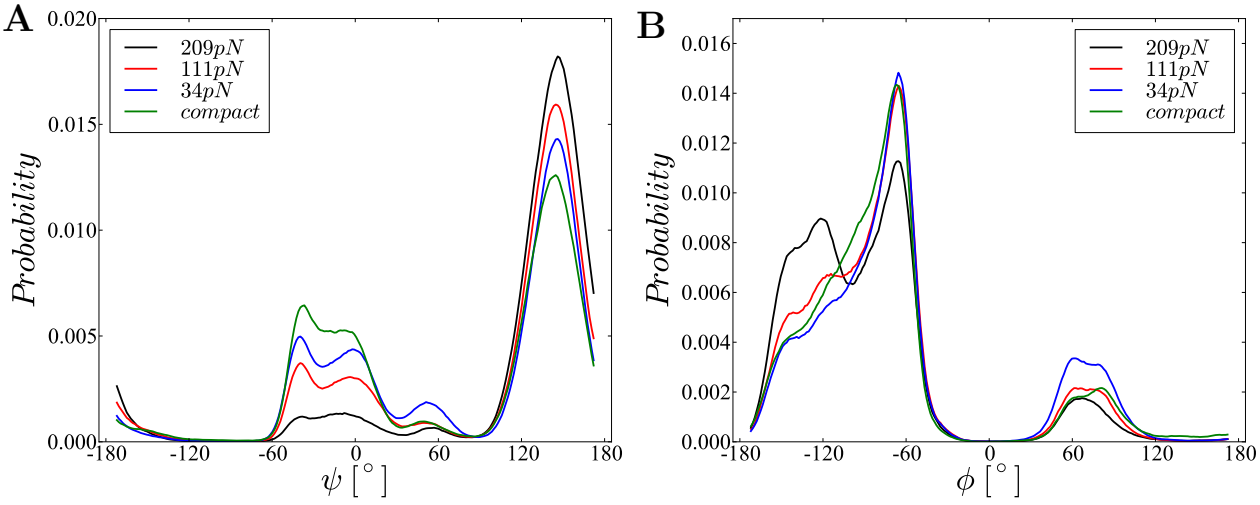


Figure 2.5: Distribution of dihedral angles for the wheat-gluten mimetic polypeptide in the compact and stretched states. (A) Distribution of ψ dihedral angles averaged over all residues. (B) Distribution of ϕ dihedral angles averaged over all residues except proline. Statistical analysis was made using simulation data in the last 9 ns, corresponding to steady-state conditions.

At intermediate tension,111 pN, dihedral angles explored smaller values and the turn configurations re-appeared ($\sim$ 20% of the fully coiled structure at 209 pN tension was converted to turn/helical conformations). At this intermediate tension, the elasticity of the

chain mainly arises from coil-turn transitions. Accordingly, the dihedral angles distribution changed noticeably from 209 pN to 111 pN. In addition, intramolecular interactions, such as hydrophobic contacts and hydrogen bonds (Table 2.3), begin emerge at this tension. As a result, transient intramolecular interactions start to affect the local elasticity of the chain, leading to fluctuations in response that may reduce the validity of the WLC model at this and lower tensions, due to the progressive breakdown of the mean field approximation that is an implicit principal assumption in polymer models such as WLC model.

The impact of amino acid sequence heterogeneity was further investigated using repeat motifs in the WGM polypeptides. Three motifs (PTSL, PGQG, and PEYD) were selected and their confirmations were studied at different forces. For each motif, propensities for the secondary structures were calculated and averaged over their repeats along the chain. These results are presented in Table 2.4. The propensity for turn structures in the PTSL peptide motif was twice that of the other two motifs, perhaps due to presence of hydrophobic residues at the ends of this motif, which make turn configuration more energetically preferable. Thus, the WGM polypeptide stiffness at intermediate tension, $\sim$ 111 pN, is partially controlled by coil-turn transitions which occur due to local hydrophobic contacts and hydrogen bonds.

209 pN	H/T (%)	Coil (%)
P-T-S-L	3 ± 2	97 ± 2
P-G-Q-G	0.4 ± 0.3	99.5 ± 0.5
P-E-Y-D	0.2 ± 0.1	99.5 ± 0.1

111 pN	H/T (%)	Coil (%)
P-T-S-L	33 ± 1	67 ± 1
P-G-Q-G	18 ± 4	82 ± 4
P-E-Y-D	14 ± 3	86 ± 3

35 pN	H/T (%)	Coil (%)
P-T-S-L	59 ± 4	41 ± 4
P-G-Q-G	52 ± 6	48 ± 6
P-E-Y-D	26 ± 3	74 ± 3

compact	H/T (%)	Coil (%)
P-T-S-L	64 ± 2	36 ± 2
P-G-Q-G	77 ± 5	23 ± 5
P-E-Y-D	28 ± 2	62 ± 2

Table 2.4: Structure propensity of three short motifs from the wheat-gluten mimetic polypeptide in compact and stretched states. Motif sequences are given using the one-letter amino acid code (P:Proline; T:Threonine; S:Serine; L:Leucine; G:Glycine; Q:Glutamine; E:Glutamic acid; Y:Tyrosine; D:aspartic acid). The average conformation in helical (α-, π-, 3_{10}-helix) or turn conformations and in coil conformation is tabulated. Statistical analysis was made using simulation data in the last 9 ns, corresponding to steady-state conditions, that has been divided into 3 batches of 3 ns each to calculate averages and standard deviations.

At low tension, 35 pN, $\sim$ 25% of coil structure at 111 pN has been converted to turn/helical conformations (Table 2.3), whereas the corresponding dihedral angles distributions were not

significantly changed (Figure 2.5). On the other hand, the number of hydrophobic contacts and backbone hydrogen bonding event were roughly doubled upon decreasing tension from 111 pN to 35 pN, implying an increasing role of local interaction on the elastic response at this relatively weak tension. This was accompanied by a significant rise in in helix/turn content in the PTSL (59 ± 4%), PGQG (52 ± 6%), and PEYD motifs (26 ± 3%), at the expense of coil content as the tension was decreased to 35 pN. However, note that for the PEYD motif in particular, high coil content (74 ± 3%) remains.

Upon further reduction of tension, the number of hydrophobic contacts and backbone hydrogen bonding event were increased three-fold from 35 pN to zero tension (the compact structure). At such low tension, intermolecular interactions play the dominant role in polypeptide elasticity (as it is noted in the number of native contacts between the sidechains), and the polypeptide no longer deforms as a classical worm like chain, but rather chain extension proceeds through a series of ruptures of transient sidechain contacts in a manner that is highly non-reversible. Indeed, hydrophobic interactions have a great impact on the global configuration of the polypeptide at low tension, bringing turns close together to form a hydrophobic polypeptide core.

Overall, the elasticity of the WGM polypeptide can be attributed to different mechanisms for each force/extension regime. While in high tension/extension, dihedral angles potential exclusively contribute to the chain elasticity, at low tension this effect is negligible and transient clusters formed by favourable intramolecular interactions between sidechain play a dominant role. In intermediate force regime, local turn/coil transitions occurring via local short-lived intramolecular interactions (Hbond / HP contacts) in key motifs are partially responsible for the elasticity of the disordered polypeptide. These finding are summarized in Figure 2.6, illustrating polypeptide configurations and dominant local structure in the different tension regimes.

2.5 Concluding Remarks

In this work, structural and mechanical properties of the wheat-gluten mimetic (WGM) polypeptide were investigated employing molecular dynamics (MD) simulation of the polypeptide in explicit aqueous solution. WGM was inspired by repeat motifs in the disordered domain responsible for the distinct elasticity of native wheat gluten proteins. Therefore, this study serves as an attempt to reveal some generic molecular aspects of disordered protein

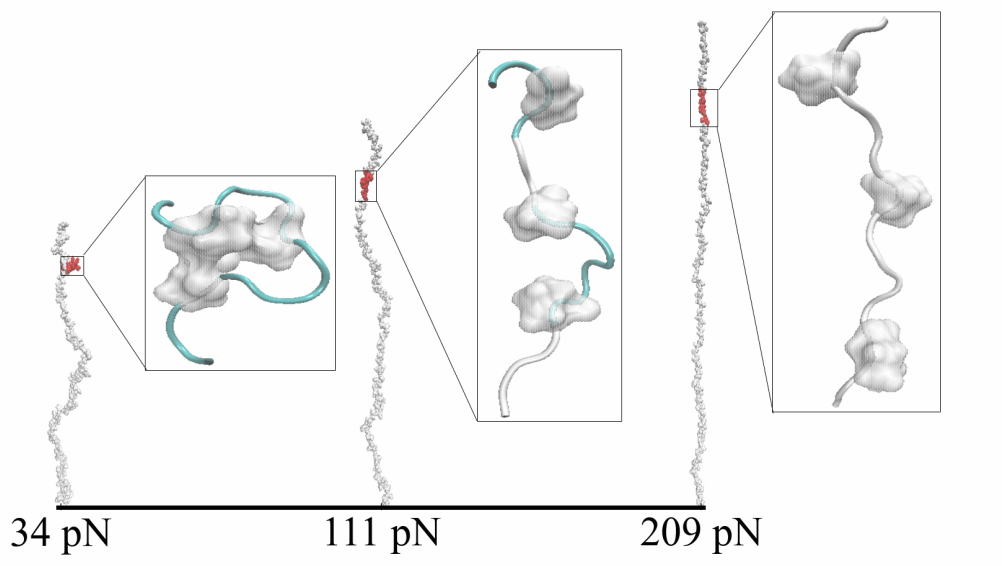


Figure 2.6: Wheat-gluten mimetic configurations and dominant local structure at different tensions. Snapshots of configurations of the WGM polypeptide and structural details of an example tandem repeat (red color) zoomed window. In this window, backbone showed in tube with white and cyan color for Coil and turn secondary structure, respectively. Also, Hydrophobic amino acid (PRO, LYS) atoms were shown in schematic VdW sphere along the snapshot of a hydrophobic core formed in low tension.

elasticity under force.

WGM showed disordered characteristics due to its high content of irregular secondary structure (turn/coil), as well as its radius of gyration which is neither as small as natively folded proteins nor as large as a chemically denatured protein of the same number of residues. On the other hand, the force-extension behaviour of WGM was assessed in series of simulations, which demonstrated distinct elastic properties compared to natively ordered proteins of the same size. The WLC model failed to predict a physical value of the persistence length and to describe local stiffness of the chain in different tension regimes, as discussed in comprehensive investigations to characterize the local mechanics of the polypeptide.

Further local structure investigation revealed two main sources for the molecular-scale elastic behaviour. The deformation of dihedral angles played a dominant role in response to strong forces. Whereas, the disruption of hydrophobic clusters played an important role in the weak force regime. The response to intermediate forces was governed by a balance between dihedral angle deformation and local turn/coil transformations in key key motifs. An analysis of the kind presented in this chapter provides a contribution to both fundamental

and applied aspects of elastomeric IDPs. Understanding the source of elastic behaviour in disordered proteins is an emerging are of interest in protein biophysics, and also is potentially useful for the design of synthetic elastic biomaterials.

2.6 Appendix

2.6.1 Modified worm-like chain model fit

At high tension, polymer elongation slightly modifies chemical bonds and contributes to the total end-to-end distance of the molecules. Accordingly, a correction was introduced by Wang et al. [57] using the Odijk [76] approach; where z/L_c in Equation 2.10 is replaced by $z/L_c - f/K$,

$$\frac{fp}{k_BT} = \frac{1}{4}\left[\left(1 - \frac{z}{L_c} + \frac{f}{K}\right)^{-2} - 1\right] + \frac{z}{L_c} - \frac{f}{K} \tag{2.11}$$

with K, the stretch modulus, ranging from 500 pN to 1600 pN in existing experimental measurements. Using above expression to fit the simulation data resulted in a p value that is still significantly smaller than peptide size, 4 Å.

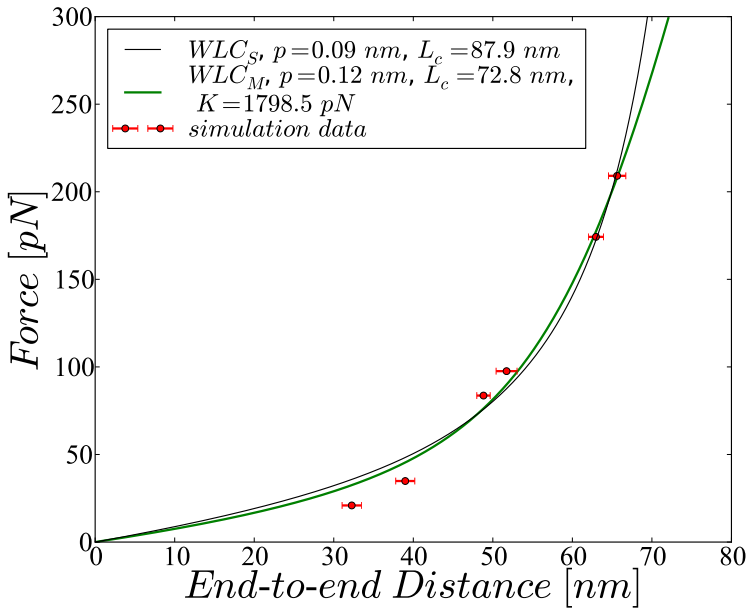


Figure 2.7: Force-extension profile along with WLC fit for wheat-gluten mimetic polypeptide. Force-extension data point extracted from simulation (red circles, average; red bars, variation), alongside two fitting curve using standard (black) and modified (green) WLC model.

2.6.2 Distribution of ϕ dihedral angles for proline

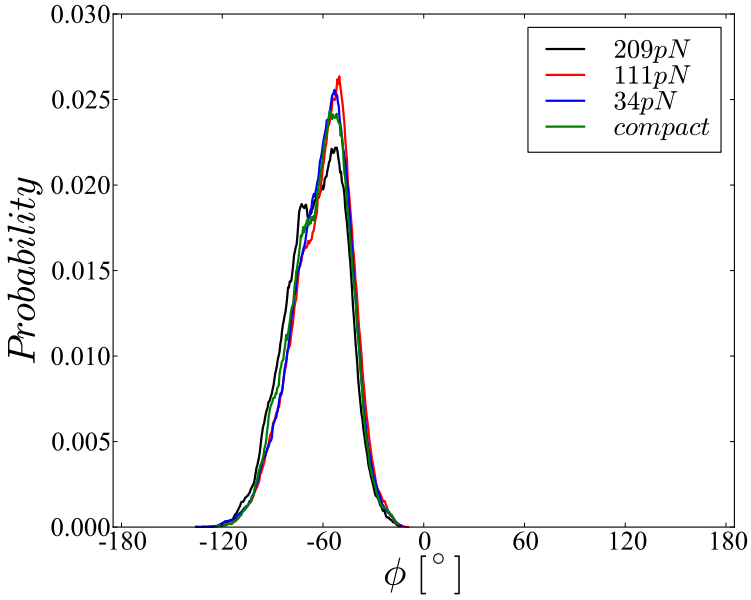


Figure 2.8: Distribution of ϕ dihedral angle for prolines residues along the wheat-gluten mimetic polypeptide under tension.

CHAPTER 3

Protein Assemblies and Functional Disorder

This work has been accepted for an oral presentation at Canadian Association of Physicists (CAP) Congress in Jun 2017, and eventually will be published in a peer-reviewed scientific journal.

3.1 Introduction

According to the widely accepted *Lock and Key* paradigm, proposed by Fisher in 1894 [16], a protein must have a well-defined 3D structure to provide a tight fit between its binding pocket and the target substrate molecule, similar to a lock fitted to a key. However, this criterion was fundamentally disputed recently through the many counterexample proteins lacking rigid three-dimensional structure, yet performimg biological functions [33]. Intrinsically disordered proteins/regions (IDPs/IDRs) are dominantly involved in cell signaling, as well as in regulating processes in which their structural heterogeneity plays a key role, such as in scaffolding and in recruitment of different binding partners [36]. While IDPs/IDRs play an important role in protein-protein interactions, little is known about the structural aspects of these interactions.

Functional IDPs/IDRs in permanent protein-protein associations have been generally classified into the two categories [77]: (i) *Assemblers*, in which disordered regions engage in protein-protein interactions as a part of the main binding site of the protein (active site), bringing multiple binding partners close together, and mediating localization of the proteins leading to higher-order protein complexes; and (ii) *Effectors*, wherein disordered regions are not included in the protein active site and do not participate in interaction directly, but modulate binding affinity of the protein indirectly, through their influence on protein conformation. Effectors have allosteric-like effects wherein they often undergo an disorder-to-order transition, known as *coupled folding*. There are many disordered assemblers and effectors involved in cellular function. For instance, RNAP II, the multi-subunit complex responsible for transcription of protein-coding genes in eukaryotes, contains a long, highly repetitive disordered assembler with a critical role in targeting and binding of messenger RNA (mRNA) to protect it from degradation [78]. On the other hand, the proline-rich (PR) domain in classical dynamin proteins acts an effector, wherein binding of this disordered domain to SH3 protein mediates dynamin self-association required for their biological pathway [79]. The function of an unstructured domain in dynamin-related protein 1(Drp1) protein, another member of dynamin superfamily, is the main scope of this study.

The dynamin superfamily are membrane remodeling enzymes which facilitate membrane remodeling events, such as the growth and scission of lipid vesicles from a membrane of one cell compartment, and the eventual fusion of the vesicle with another compartment [80]. In membrane scission reactions, dynamins act as mechanochemical enzymes that are enhanced

via GTP (guanosine triphosphate) hydrolysis activity. They lie on the target membrane where they assemble further into rings or spirals around the membrane, tightening to apply stress to the membrane and eventually inducing membrane cleavage [81]. Although dynamins engage in different membrane remodeling events, they all share similar subunit features that include a G domain which carries out the GTP hydrolysis, a stalk domain with a coiled-coil structure that enables the protein to self-assemble into linear complexes, and another domain whose function varies between dynamins. A schematic view of the single domain structure and self-oligomerization configuration for Drp1, a member of dynamin superfamily, are shown in Figure 3.1. While for some members of this superfamily (e.g., classical dynamins (dynamin I-III) it has been shown that this variable domain is responsible for associating the protein with the membrane surface, for other members (such as Drp1) the structure and function of this variable domain are still unknown.

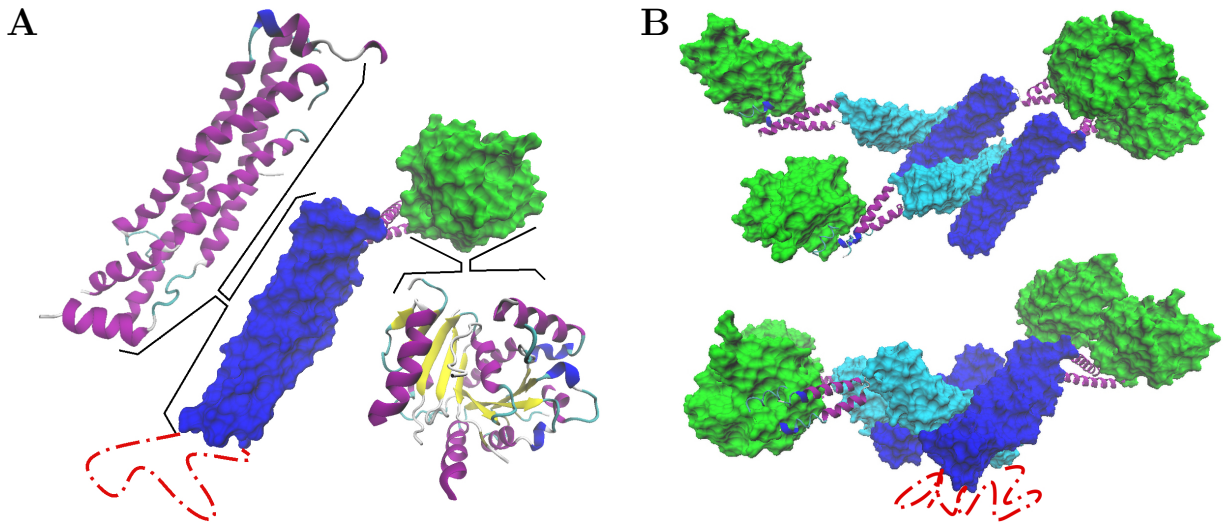


Figure 3.1: (A) Schematic view of Drp1 structure indicating its subdomains (G domain, green; Stalk domain, blue; variable domain/B domain, red dashed line) along with their secondary structure (α-helix, purple ribbon; β-sheet, yellow arrows; white/blue tube, Coil/Turn). (B) Top and side view of the protein arrangement in assembled oligomers of Drp1 (light and dark blue used for adjacent stalk domains to increase visibility). The figure was produced using X-ray crystallography results for four Drp1 bundles (PDB:4BEJ) in which the variable domain was removed to facilitate crystallization.

Our collaborators in the Hill group at the Johns Hopkins University and at the Medical College of Wisconsin have studied Drp1 and its variable domain, referred to as B domain. Initial results of these studies were presented in the Ph.D. thesis of Ammon Posey [82]. They

found that the B domain has a CD spectrum consistent with disordered proteins. It was believed that the B domain is involved in protein-membrane interactions in a similar manner to the role of the variable domain (proline-rich domain) in classical dynamins. Therefore, they proposed that the B domain behaves as a disordered effector. In their hypothesis, B domain initially acts as a spacer creating excluded volume to inhibit Drp1 from self-association (an entropic bristle barrier), which would be eliminated by folding to a compact structure upon the exposure to favourable environmental conditions (a coupled-folding mechanism). In support of their hypothesis, the Hill group was aiming to show that folding of B domain will facilitate Drp1 assembly. To induce folding, they used trimethylamine N-oxide (TMAO) osmolyte, which is known to increase the thermodynamic stability of folded proteins in harsh conditions [83] and also has been shown to cause folding in some disordered proteins [84]. These osmolytes are able to create a crowded environment that entropically stabilizes protein structure in some cases [85]. Experiments on isolated B domains in TMAO solution indicated that the domains were significantly involved in self-interactions resulting in protein coacervation. Although TMAO clearly affects B domain interactions, the mechanism of TMAO-induced association was not clear. In particular, CD studies did not indicate a substantial change in B-domain secondary structure in the presence of TMAO, thus raising doubts about a potential coupled-folding pathway.

The focus of this simulation study is on the B domain structure and its role in protein self-assembly in solution. Here, the degree of disorder of the B domain conformation in aqueous solution, as well as possible structural transitions in presence of TMAO were investigated. Moreover, B domain self-association was studied to examine its contrasting proposed functions as putative entropic bristles prohibiting self-assembly and alternatively as a promoter of the Drp1 assembly process via specific inter-domain self-interactions. To achieve this latter goal, numerous potential self-association configurations of B domain pairs were constructed, equilibrated, and the corresponding binding energy calculated for stable dimers in two different solutions (water-NaCl and water-TMAO solution).

3.2 Simulation and Analysis Methods

3.2.1 General simulation details

A series of simulations were performed in which conformation of the B domain in different solutions was monitored and possible dimer configurations and their related binding energies were investigated. In all simulations, peptide chains were solvated within neutral water-ion solution (0.2 M NaCl) with periodic boundary conditions using the *Solvate* and *Autoionize* module of VMD [52]. To add TMAO osmolytes to the solutions, these molecules were placed in a uniform cubic lattice configuration within a simulation box, with the lattice unit size chosen to provide the desired TMAO solution concentration. Subsequently, those TMAO molecules within 2.5 Å from the solvated peptide and water/ion molecules within 1.5 Å from TMAO/peptide were removed. The simulation box was extended to maintain a minimum of 15 Å between the peptide atoms and the boundary of the simulation box. The final TMAO solutions for single and dimer B domain simulations contained 3.3 M and 5.1 M TMAO, respectively. TMAO molarity was assigned according to the ratio of total water/TMAO molecules in a given solution, with a defined molarity of 55.5 for water.

All molecular dynamics (MD) simulations employed a Langevin thermostat and barostat and were conducted in neutral water-ion solutions in an isothermal-isobaric ensemble (NPT) at 300 K and 1 bar. The CHARMM36 force field [68] with the protein backbone improvement [69] and dihedral angles optimization [86] was adopted for the peptides and the TIP3P model was used for the water molecules [70]. Force field parameters for neutral TMAO molecules were obtained from the modified version of Kast et al. model [87], which are optimized by Garcia et al [88] to achieve correct preferential exclusion behavior for TMAO molecules. A cut-off distance of 12 Å was used for electrostatic and Lennard-Jones interactions and electrostatic interactions were computed via the particle-mesh Ewald (PME) algorithm. All atomistic simulations were carried out using NAMD 2.10 [71] with a time step of 2 fs.

3.2.2 B domain equilibration

The B domain sequence contains the following 137 amino acid residues:

```
FADACGLMNN-NIEEQRRNRL-ARELPSAVSR-DKSSKVPSAL-PASQEPSPA-ASAEADGKLI-...
```

```
...-QDSRRETKNV-ASGGGGVGDG-VQEPTTGNWR-GMLKTSKAEE-LLAEEKSKPI-PIMPASPQKG-...
...-HAVNLLDVPV-PVARKLS
```

Using the above sequence a polypeptide chain was constructed in a random coil configuration, for which there is no native contact or secondary structure, using a Python code developed by the author (See Appendix B). Subsequently, this random coil was solvated in a simulation box of size (72Å x72Å x72Å), and equilibrated over 300 ns in aqueous buffer and TMAO solutions. Data from the last 50 ns of simulation, for which the polypeptide reached a stable, steady-state structure (as determined by the small fluctuations in the radius of gyration of the polypeptide), was used for single domain analysis. Furthermore, the dimerization process was stimulated for both solution conditions using two distinct, randomly selected B domain conformations from last 50 ns of the single domain simulation. These two B domain configurations were initially separated by a surface-to-surface separation of 20 Å and simulated for multiple runs 200 ns to allow for the possible formation of a dimer structure. Then, final configuration of the proteins system were simulated in smaller simulation box for another 200 ns.

3.2.3 Steered MD to construct various dimer configurations

The equilibrated dimer structures were further employed to explore possible intermolecular associations leading to oligomerization of the B domain. To achieve this goal, six orthogonal binding directions, analogous to Cartesian coordinates ±X, ±Y, and ±Z, were initially tested for each B-domain in an equilibrated dimer configuration. For example, a dimer defined by the $+Z_1$ site of molecule 1 that faces the $+Z_2$ site of other molecule 2, as shown in Figure 3.4, is referred to here as $+Z_1+Z_2$. Starting such an equilibrated $+Z_1+Z_2$ dimer, the individual B-domains were rotated, while their inter-molecular separation was maintained, to construct all 36 test association configurations between two B-domains for each solution condition. These potential dimers were further equilibrated to relax the initially imposed contact surfaces and thereby search the local orientation space for stable associations. Simulations were conducted for 25 ns in each relative pair orientation, for a total of 1800 ns simulation time. A harmonic potential between the centers-of-mass of the B-domains with a force constant of 0.2 kCal/mol.$Å^2$ was applied to bias B-domains pairs against diffusing apart during the dimerization search process.

Among all 72 conformations for both water and TMAO solutions, the most promising

candidates for stable association were selected based on two criteria: the contact surface of binding sites and the force required to hold the molecules together. Surface contact was defined by the number of atoms within 3 Å of both molecules and the force was estimated from the separation of the centers of mass, using the harmonic potential spring constant. The contact surface criterion was determined from the steady-state value that was obtained for the initial stable dimer (the $+Z_1+Z_2$ dimer) discovered in extended, constraint-free simulations of two B-domains for 400 ns. Dimer structures that showed surface contact $> 80\%$ of surface contact in the $+Z_1+Z_2$ reference dimer and have an intermolecular attractive force or a repulsive force smaller than 0.53 kCal/molÅ were considered as potential stable association states. Overall four such configurations for water solutions and eight ones for TMAO solutions were selected for binding energy measurement.

3.2.4 Binding energy calculation for dimer

Assignment of energy to a state of association requires a comprehensive sampling of the system as a function of appropriate reaction coordinates. The existence and identity of such reaction coordinates for complex molecules such a polypeptide are not known a priori, but must be postulated and confirmed empirically. Moreover, such a sampling of the system can be challenging, especially when a rare event is desired in which simulation time scales are shorter than event occurrence timescales, such often occurs for protein-protein association/dissociation.

In this case, force-biased molecular dynamics simulation is commonly used to accelerate rare events. Umbrella sampling (US) [89] was used here to find the binding energy for the selected B-domain associations. In this method, a force bias is applied to the system and moves it along the reaction coordinate in multiple time intervals, referred to as windows. In each window, the system configuration is equilibrated and sampled while the Boltzmann weighting is adjusted according to the applied auxiliary force in order to properly weight observed samples. Appropriate set-up in terms of applied force, sampling period, and window size should provide sufficient overlap between adjacent windows to insure a continuous pathway between the initial and the final system configuration. Integrating results from all windows then provides the global free energy evolution along the reaction coordinate.

To perform umbrella sampling for the dissociation process of a dimer complex, the distance between the center of mass (COM) of the molecules were considered as the reaction coordinate

and the COM was calculated from the residue C_α coordinates. To manipulate dimer structure and eventually dissociate the molecules, the value of this reaction coordinate was harmonically restrained to variable d_v, which was increased sequentially throughout the simulation. A harmonic potential with a force constant of 2.5 kCal/mol.Å^2 was used to apply restraint and couple the reaction coordinate to d_v value, which was increased in 36 and 48 windows, 0.5 Å each, for water and TMAO solution conditions, receptively. The proteins complexes were equilibrated and sampled for 0.6 ns for each interval. During this process, a second harmonic potential with a force constant of 5 kCal/mol.Å^2 was applied to individual peptide backbone molecules to approximately preserve their initial equilibrated configurations in the dimer assembly while undergoing the dissociation process. The backbone root mean square displacement (RMSD) was used as the potential variable for this constraint.

The sampled data obtain through this dissociation process was further analyzed using the weighted histogram analysis method (WHAM) [90] to adjust Boltzmann weighting accordingly and obtained the energy change during the dissociation process while considering the contribution of both constraints. This task was accomplished by `wham` code developed by Grossfield lab. The variance in the free energy was estimated by the square of the cumulative statistical error in the estimate of the mean position of reaction coordinates [91].

3.2.5 Description of analysis and methods

All measurement were performed when the polypeptide showed sufficiently small structural fluctuations ($R_g < 0.5$ Å), indicating that the molecules had reached a steady state average configuration appropriate for further analysis. The last 50 ns of simulations of the B domain in either single and dimer arrangements, in presence and absence of TMAO osmolyte molecules, were divided into the five equal batches of 10 ns each for statistical analysis. The coordinates of backbone alpha-carbons (C_α) were used for radii of gyration calculations.

Secondary structure content was measured in three categories: Helical, Turn, and Coil using the STRIDE algorithm [72] implemented in VMD [52]. The beta content was found to be negligible and so was not reported. To obtain solvent distribution, the number of water and TMAO molecules within a given distance from the polypeptide chain were calculated using coordinates of water oxygen and TMAO nitrogen. The ideal water/TMAO ratio was defined as the total number of water molecules over the total number of TMAO molecules in a given solution.

Native contacts (peptide-peptide, peptide-solvent, solvent-solvent, solvent-solute) in all simulations were quantified in terms of hydrogen bonds between donor and acceptor atoms. This also accounts for sidechain charged group (amine and carboxyl) in the polypeptide sequence. A hydrogen bond was thus defined by the criteria of separation and orientation of polar *Donor* and *Acceptor* atoms. A hydrogen bond, of the form Donor-Hydrogen...Acceptor, was noted when the distance between donor and acceptor is less than 3.2 Åand its bond angle, defined in terms of the deviation of Donor-Hydrogen-Acceptor from linearity, is less than 40°. Thus, an ideal bond angle is 0° based on the bond angle definition here.

Native contacts such as protein-protein, protein-solvent, solvent-solvent in all simulations are quantified in term of hydrogen bond that also accounts for electrostatic interactions between opposite charge residues. A hydrogen bond was defined by criteria of special separation and orientation of polar atoms as *Donor* and *Acceptor*. A hydrogen bond, Donor-Hydrogen...Acceptor, is formed when the distance between donor and acceptor is less than 3.2 Åand its angle is less than 40°. Hydrogen bond angle is defined as the deviation of Donor-Hydrogen-Acceptor from linearity; thus, an ideal bond angle is 0° based on the bond angle definition here.

The distance map between two domains for a given dimer, as defined as the shortest distance between any sidechain atom of a pair amino acids, was represented as a two-dimensional heat-map diagram using the distance between inter-chain residues pairs averaged over the last 10 ns after the establishment of steady state conditions.

3.3 Results and Discussions

3.3.1 B domain has disordered structure

The proposed hypothesis (coupled-folding) for a structural transition in the B domain protein was addressed through series of simulations including stabilizer osmolytes and dimerization. The B domain was equilibrated over 300 ns in explicit water-ion solutions starting from an extended random coil structure where there is no native intramolecular contact. In a similar way, the B domain was also simulated in the presence of trimethylamine N-oxide (TMAO), a small neutral but polar osmolyte molecule. A snapshot of the final compact configuration along with its secondary structure details and charge residue associations were shown in Figure 3.2. Moreover, a possible couple folding and structural transition via self-association

of the B domain was investigated. Two compact configurations from the steady state of the previous single B-domain simulation were selected for each solution and placed apart from each other in a simulation box (with and without TMAO molecules, according to their original solution conditions) and simulated over 400 ns, a period sufficient to form stable dimer complexes in both solutions. Radii of gyration and secondary structure content were measured in the final compact steady-state structure of the B domain both in the presence and absence of TMAO osmolyte and for both monomer and dimers simulation. The results are presented in Table 3.1. Note that statistical analysis of the data was performed on conformation in last 50 ns of all simulations, where the radius of gyration had reached a plateau and its fluctuation was small, ensuring that results represented molecular characteristics in steady-state conditions.

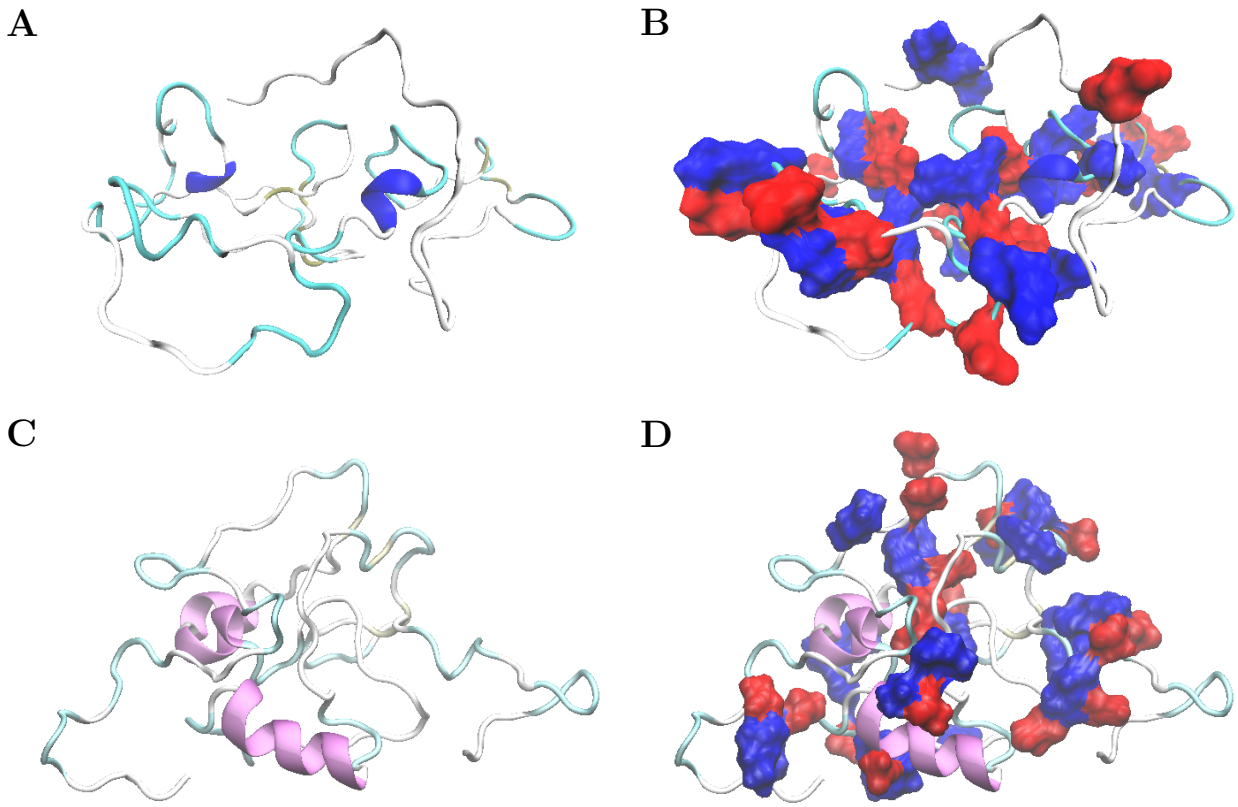


Figure 3.2: Snapshot of the compact configuration of the B domain in water and TMAO solutions.The secondary structure of the protein backbone (white tube, Coil; cyan tube, Turns; purple ribbon, α-helix) was determined using STRIDE methods [72] for water (A) and TMAO (C) solution. A colour map of electrostatic interactions between positive (blue) and negative (red) residues for water (B) and TMAO solution (D).

We first examine the global conformational characteristics the B-domains. Isolated B-domains in solution form globular chain conformations (see Figure 3.2) that can be characterized by a radius of gyration R_g. In the absence of TMAO in solution, the simulations indicated R_g=17.2 ± 0.3 Å, while in the presence of TMAO, R_g=16.5 ± 0.3 Å. Thus, contrary to expectations for the effect of an osmolyte in the solution, the presence of TMAO was found to have no significant effect on the B-domain radius of gyration.

To gain further insight into the global B-domain chain conformation, we employed a scaling analysis used in polymer physics [92] to relate the radius of gyration, R_g to the number of amino acids in the B-domain, N, via $R_g = bN^{\nu}$, where b is a characteristic size of an amino acid residue and ν is a scaling exponent indicative of the degree of compactness of a polymer chain. Recall from Chapter 2 that for a simple homopolymer chain of N monomers of size b, $\nu \simeq 1/3$ for a compact globule, $\nu \simeq 1/2$ for an entropic random walk conformation, and $\nu \simeq 3/5$ for a expanded self-avoiding walk conformation. Proteins have been classified using such a scaling approach, wherein compact (possibly folded) globular proteins have $\nu \simeq 1/3$ and denatured proteins have coil-like characteristics with $\nu \geq 1/2$ [73,93]. Using the values for the monomeric B-domain radius of gyration in Table 3.1 and typical values for b (2.0 ± 0.2 Å [73,93]), the apparent exponent for this domain was found to be $\nu = 0.44 \pm 0.04$, independent of added TMAO osmolyte. This monomer B-domain exponent lies between the values for natively folded ($\nu = 0.33$) and denatured ($\nu \geq 1/2$) proteins, consistent with a disordered, weakly solvated, molten globule conformation. Interestingly, R_g was found to slightly decrease upon dimer formation (discussed in more detail below) while still insensitive to the presence of TMAO in solutions ($15.5 \leq R_g \leq 16.1$ Å, with a corresponding exponent $\nu \simeq 0.42$, consistent with a slightly more compact conformation.

We next discuss the internal secondary structure of the B-domains. A single, isolated B domain in 0.2 M NaCl aqueous solution showed only 2% helical content, the rest being irregular secondary structure. The addition of TMAO osmolyte to the solution increased the helical content of the B domain to 14%, a rather modest effect. Thus majority of the B domain is maintained an irregular secondary structure for both solution conditions. Consequently, we can conclude that TMAO osmolyte failed to induce a noticeable structural transition in isolated B domains, which remained largely in a disordered globular structure. Moreover, analogous simulations on B-domain dimers in water and TMAO solutions (discussed in more detail below) indicated a only very modest increase helical content (up to 16%) for both solution conditions. These results are in good agreement with CD results observed by our

collaborators [82], indicating highly disordered structure for the B domain protein, even when crowded by an osmolyte.

Overall, since neither the presence of TMAO osmolyte nor dimerization of the B-domains had a significant effect on the glocal conformation and the internal structure of the B-domain, we can conclude that a coupled-folding process is not operative in the B-domain as an independent polypeptide. One possibility that cannot be ruled out is that the behaviour of B-domain is qualitatively different as a integral part of the native Drp1 protein. Another possibility is that the function of the B-domain in Drp1 is to modulate Drp1 self-assembly, but perhaps not in a strictly inhibitory modality (as would be the case for a classic entropic bristle). This second possibility is explored in the remaining part of this chapter.

	R_g (Å)	Helical (%)	Turn (%)	Coil (%)
Monomer	17.2 ± 0.3 [16.5 ± 0.3]	2 ± 0 [14 ± 0]	43 ± 0 [40 ± 0]	50 ± 0 [42 ± 0]
Dimer: chain 1	15.6 ± 0.1 [15.6 ± 0.1]	12 ± 2 [16 ± 1]	40 ± 5 [40 ± 1]	44 ± 2 [43 ± 1]
Dimer: chain 2	16.1 ± 0.2 [15.9 ± 0.2]	16 ± 2 [17 ± 1]	38 ± 6 [34 ± 1]	41 ± 3 [43 ± 1]

Table 3.1: Local structure and global size of the B domain protein. Average radii of gyration (R_g) were measured for the B domain in monomer and dimer simulations. In addition, the average content of secondary structures for all conditions was calculated using STRIDE methods [72]. Beta content was not shown due to its insignificant amount. Data is provided for B-domains in water and TMAO solutions (the latter shown in square brackets). Statistical analysis was made using simulation data in the last 50 ns, corresponding to steady-state conditions, that has been divided into 5 batches of 10 ns each to calculate averages and standard deviations.

3.3.2 Molecular interaction of TMAO with B-domains

Our experimental collaborators noticed that B domain molecules aggregate to form microscopic blobs upon addition of TMAO osmolyte to the B domain solution. In contrast to their hypothesis, no significant structural transition was induced in B domains via TMAO osmolyte, as indicated in the previous section. Therefore, another mechanism must be employed by TMAO osmolyte to enhance B domain association. TMAO is mainly known as a protective osmolyte which is depleted from the vicinity of the protein backbone, thereby enhancing the folding process by modifying the hydration of the protein. To investigate the intereaction of TMAO with the B-domain, the solvent composition near polypeptide backbones at the solvated surface of the B-domain globule was evaluated by measuring the

average number of TMAO and water molecules in radial shells of thickness 4 Å. Analyses were performed for both single and dimer simulation data, and results are given in Table 3.2 in terms of the fraction of water to TMAO molecules.

Within the first solvation shell of a single, isolated B domain, there was a significant increase in the ratio of water to TMAO molecules as compared with the ratio in homogeneous bulk solution (the total water in the solution over the total TMAO molecules). However, the water to TMAO ratio was below the ideal ratio in the second solvation layer. Beyond the second shell, the ideal bulk solution value of this ratio was recovered. Similar behavior was observed for B domains in dimeric structures. Therefore, we may conclude that TMAO molecules were preferentially excluded from the first solvation layer but enhanced in the second layer, in agreement with the definition of a protective osmolyte. This exclusion and inclusion should reflect the nature of backbone and sidechain interactions, respectively.

B domain: Monomer			
Distance	Avg # of Waters	Avg # of TMAO	Water: TMAO
< 4 Å	642 ± 11	28 ± 2	22.9:1
4–8 Å	736 ± 16	48 ± 2	15.3:1
8–12 Å	792 ± 18	46 ± 1	17.2:1
12–16 Å	840 ± 20	50 ± 2	16.8:1

B domain: Dimer			
Distance	Avg # of Waters	Avg # of TMAO	Water: TMAO
< 4 Å	706 ± 6	51 ± 1	13.9:1
4–8 Å	988 ± 6	97 ± 1	10.2:1
8–12 Å	1247 ± 15	113 ± 1	11.0:1
12–16 Å	1537 ± 26	140 ± 4	11.0:1

Table 3.2: Solvent distribution in radial shells measured from the B domain surface. The number of water oxygens and TMAO nitrogens found in shells within a certain distance from the B domain surface residue backbone is listed separately in monomer and dimer simulations. Statistical analysis was made using simulation data in the last 50 ns, corresponding to steady-state conditions, that has been divided into 5 batches of 10 ns each to calculate averages and standard deviations. The bulk water:TMAO ratio is 17:1 for monomer solutions and 11:1 for dimer solutions.

TMAO-polypeptide interactions were further analyzed by investigating the average number of peptide-water, peptide-TMAO, and peptide-peptide hydrogen bonds. Results for both monomer and dimer simulations are reported in Table 3.3. There was a significant decrease in the average number of water-peptide hydrogen bonds in TMAO solution as compared with water solution. Since TMAO molecules were partially excluded from the vicinity of the protein backbone, this reduction was mainly related to the peptide sidechain interface with solvents where TMAO molecules contacts compete with and interrupt water-sidechain interactions. The presence of TMAOs, however, was not able to improve the number of

intramolecular hydrogen bonds within the B domain and thus failed to fold the protein. On the other hand, there was a significant interaction between TMAOs and peptide sidechains on the B domain surface, where 15 peptide-TMAO contacts for the single domain and 39 contacts for the dimer structure were observed. This suggests that the main effect of TMAO was to interrupt sidechain-water hydrogen bond on the surface of the B-domain and in the gap between two B-domains in a dimer configuration.

B domain: Monomer			
Solution	Water-Protein	Protein-Protein	TMAO-Protein
Water	334 ± 13	54 ± 5	—
TMAO	297 ± 12	59 ± 6	15 ± 3

B domain: Dimer			
Solution	Water-Protein	Protein-Protein	TMAO-Protein
Water	646 ± 17	119 ± 6	—
TMAO	561 ± 15	120 ± 5	39 ± 5

Table 3.3: Average number of hydrogen bonds for a B domain in a monomer and dimer configuration. Statistical analysis was made using simulation data in the last 50 ns, corresponding to steady-state conditions, that has been divided into 5 batches of 10 ns each to calculate averages and standard deviations.

TMAO is a neutral osmolyte but it has a strong dipole moment. Its oxygen can act as a hydrogen bond acceptor and enable it to interacts with the backbone, as well as with charged and polar sidechains. To investigate dominant TMAO interactions, protein residues which were in contact with TMAO osmolytes for more than 10% of the simulation time were recorded, along with their total occupancy with TMAO molecules. Figure 3.3 summarizes this data. Residues with a positively charged sidechain group, such as Lysine and Arginine, dominantly interacted with TMAO and even were able to form a hydrogen bond with two TMAO molecules simultaneously, resulting in an apparent occupancy of more than 100%. Although uncharged polar residues such as Serine, Glutamine, Threonine, and Asparagine also formed hydrogen bonds with TMAO molecules, they were not as frequent or stable in their association as the residues with positive residues. This suggests that uncharged polar residue contacts with TMAO molecules are energetically comparable to their contact with water molecules. However, TMAO forms a more favourable bond with positively charged sidechain than with water molecules. Therefore, despite the strong excluded volume influence of TMAO near the B-domain surface, these osmolytes were nevertheless able to interact with positive residues directly (evidence which was also observed in [94]). These persistent interactions between TMAO and positively charged residues possibly reduce intermolecular electrostatic repulsion and thereby facilitate inter-chain associations leading to aggregation/assemblies of

B domain that have been experimentally observed by our collaborators.

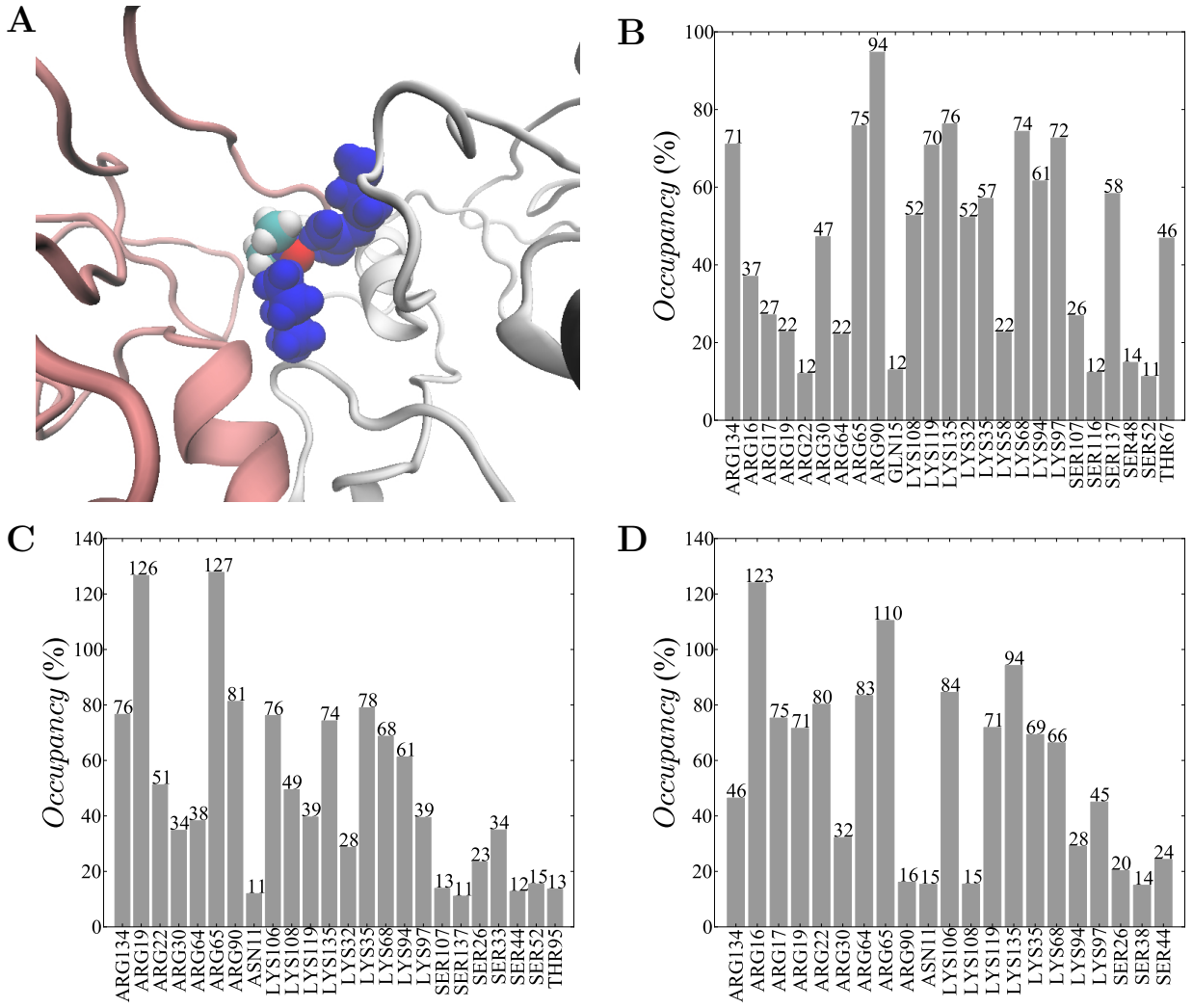


Figure 3.3: (A) Snapshot of TMAO oxygen interactions with two positive charges in an inter-chain association pocket (chain 1, white; chain 2, pink; positive charged residue, blue; TMAO oxygen , red). Occupancy of B domain residues by TMAO osmolyte in single domain simulation (B) and in dimer simulation for chain 1 (C) and chain 2 (D). The criterion for a contact event was that residues have occupancy more than 10%, and occupancy percentage was calculated based on accumulated residence time for an individual TMAO contact with given residue. Note that residues able to bind to more than one TMAO simultaneously can correspond to more than 100% occupancy.

Interestingly, positively charged amino acids within the binding pocket of a dimer structure (e.g., Chain1: 64Arg, 65Arg, 119LYS; Chain2: 90Arg, 94LYS, 97 LYS) showed noticeable association with the TMAO osmolytes. In such a scenario, two B domains were able to form

a loose dimer in which a TMAO molecule inside the pocket can transiently interact with individual positively charged residues, thereby reduce the energy cost of having like-charged residues in the binding pocket. In addition, TMAO osmolytes were occasionally able to form an intermolecular bridge by binding simultaneously to two positive residues on the chain surfaces, thereby facilitating dimerization (Figure 3.3 A). These results suggest that TMAO molecules reduce the electrostatic repulsion between the B domains, in both direct and indirect way, and as a result mediate dimer formation. This concept can be further tested by a systematic examination of intermolecular association between two B-domains in different relative orientations.

3.3.3 Inter-chain associations of the B domain

In presence of the TMAO osmolyte, B domains form coacervates (an indication of higher order assemblies), whereas this phenomenon was not observed in pure water solution. Therefore, it is likely that TMAO molecules facilitate higher order associations of the B domain, which would required to induce self-oligomerization higher than a dimer. To characterize the association process, a systematic exploration of the contact surface between two B domains was performed for pairs of B-domains in both water and TMAO solutions. Six orthogonal potential associating surfaces on each molecule were considered by defining an orthonormal coordinate system on each B-domain, wherein the $+Z_n$ direction for $n = 1, 2$ was chosen from the direction from the centre of mass to a site of strong association on molecule 1 and 2, respectively, that was found in an initial unbiased 400 ns simulation of dimer formation (See Figure 3.4). Thus, the surface each B-domain in a putative dimer pair was assigned 6 distinct regions or "faces", denoted $\pm X_1$, $\pm Y_1$, $\pm Z_1$ for B-domain 1 and $\pm X_2$, $\pm Y_2$, $\pm Z_1$ for B-domain 2, to define 36 initial conditions, using all different pair combinations of these faces, to explore possible dimer formation.

Eight association configurations for pairs of B-domains were found for the TMAO solutions and four ones were found for the water solution were identified as stable dimer candidates, based on their higher stability and elevated intermolecular contact area. For these, binding energies were estimated by measuring the energy required to dissociate them, using the umbrella sampling (US) methods discussed earlier in this thesis. The results are presented in Table 3.4, which show binding energy values corresponding to these candidate dimer structures for B-domains in water and in TMAO solutions. Those candidate dimer entries shaded

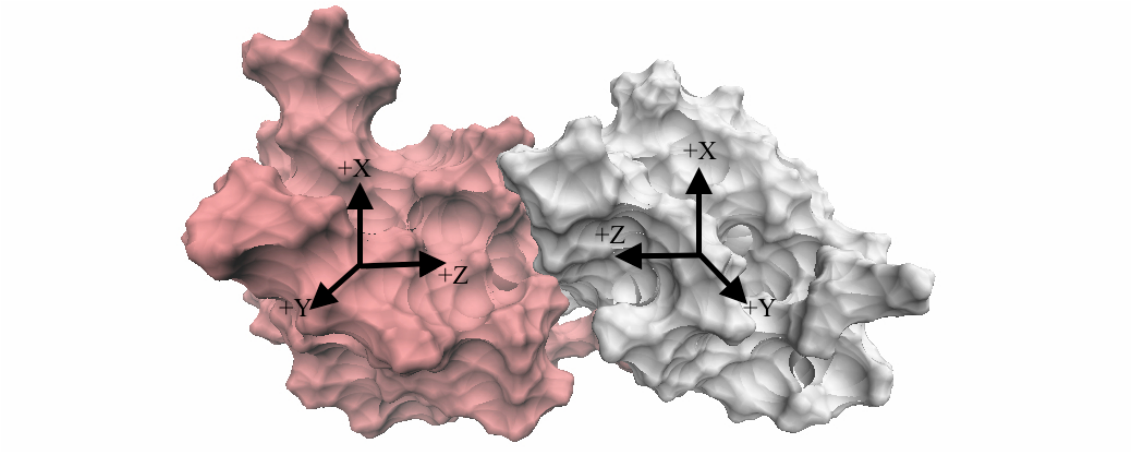


Figure 3.4: A schematic of dimer structure and reference coordinates on each domain (chain 1, white; chain 2, pink).

in gray correspond to cases where the magnitude of the binding energy exceeds a threshold criterion (values corresponding to found stable dimer structure +Z+Z pocket with 21.0 ± 0.5 kCal/mol for the water solutions and 16.1 ± 0.4 kCal/mol for the TMAO solutions) and thus were considered to be stable dimers. However, the entropy lost when two independent B-domains form a dimer was not considered in binding energy calculation; therefore, dimer with binding energy less than these thresholds do not necessarily have stable structure. In accordance with this criterion, we identified four stable dimer for the TMAO solutions and one for the water solution. We hypothesize that higher order aggregates, including macroscopic coacervates, could form from B-domains in the TMAO solutions, since each B-domain could in principle bind four other B-domains. This is consistent with experimental evidence indicating B domain coacervates form in TMAO solutions but not in water.

The binding pockets suggested by the simulation results were further analyzed in terms of the residues involved in interchain association, as characterized by a distance map between residues of the two chains determined by averaging over MD trajectories. Figure 3.5 shows such a distance map for the case of the stable $+Z_1+Z_2$ dimer in water. In this case, there are about 30 pairs of residues which were within 4 Å of each other, indicating strong sidechain overlap and the significant contact surface between the two B domains in water. In particular, Figure 3.5 clearly indicates the notable contribution of hydrophobic residues such as Proline, Valine, Leucine, Methionine, and Isoleucine in the contact region (for instance 129PRO–PRO84, 129PRO–PRO109, 132VAL–ILE110, 50ALA–ILE110). The hydrophobic interaction appears to be the main driving force for dimer formation in the water solution.

Binding Pocket	Binding Energy(kCal/mol)
$-Y_1^t+Z_2^t$	-23.9 ± 0.3
$-Y_1^t+Y_2^t$	-22.9 ± 0.4
$+Y_1^t+X_2^t$	-21.8 ± 0.3
$+Z_1^t+Z_2^t$	-16.1 ± 0.4
$-Y_1^t-Z_2^t$	-13.3 ± 0.4
$-Y_1^t-Y_2^t$	-11.4 ± 0.3
$-Y_1^t+X_2^t$	-10.6 ± 0.3
$+Z_1^t-Y_2^t$	-9.8 ± 0.3

Binding Pocket	Binding Energy(kCal/mol)
$+Z_1^w+Z_2^w$	-21.0 ± 0.5
$-Y_1^w-Y_2^w$	-11.7± 0.5
$+X_1^w+X_2^w$	-11.3 ± 0.6
$-X_1^w-Z_2^w$	-11.2 ± 0.5

Table 3.4: Binding energies for promising dimer configurations in TMAO (left) and in pure water (right) solutions. The most stable associations are highlighted in gray.

However, a few charged residues in the binding pocket were found to be in strong contact with opposite charges atoms in the binding partner sequence, to minimize electrostatic energy in the association pocket (for instance 127ASP–ARG134, 134ARG–GLU105[backbone O], 58LYS–ILU110[backbone O]), and thus also played a role in stabilizing this dimer configuration..

A similar analysis of inter-chain contact was performed for the four most stable dimer configurations obtained for B domains in TMAO solutions. The results are shown in Figure 3.6. In comparison to water solution, dimers in the TMAO solution showed significantly smaller contact surface, especially for -Y+Z, -Y+Y, and +Z+Z dimers (See Figure 3.6 A, B, D) for which there were about 15 pairs of residues in contact. This indicates that dimers in TMAO solution had a less tightly bound pocket, due to water and TMAO molecules intercalated in the pocket. This behaviour is consistent with the previous findings of strong association of TMAO with protein protein sidechains on the free surface of a single B-domain. Indeed, in the binding pocket between two B-domains, a TMAO molecule is able to associate with polar residues on either B-domain, thereby screening unfavourable interactions between B-domains and enhancing the stability of the dimer. Moreover, a TMAO molecule is able to simultaneously associate with polar residues on both B-domains, thereby creating a stabilizing transient bridge between B-domains. In addition, B domains in these dimer configurations were stabilized by favourable intermolecular electrostatic interactions between residues, which are often stronger than hydrophobic ones. For instance, the $-Y_1+Z_2$ and and $+Z_1+Z_2$ dimers (Figure 3.6 A and D) contains interacting ARG22, ARG90, LYS94, and LYS97 residues; while in the $+Y_1+X_2$ dimer (Figure 3.6 C) contains ARG16, ARG64,

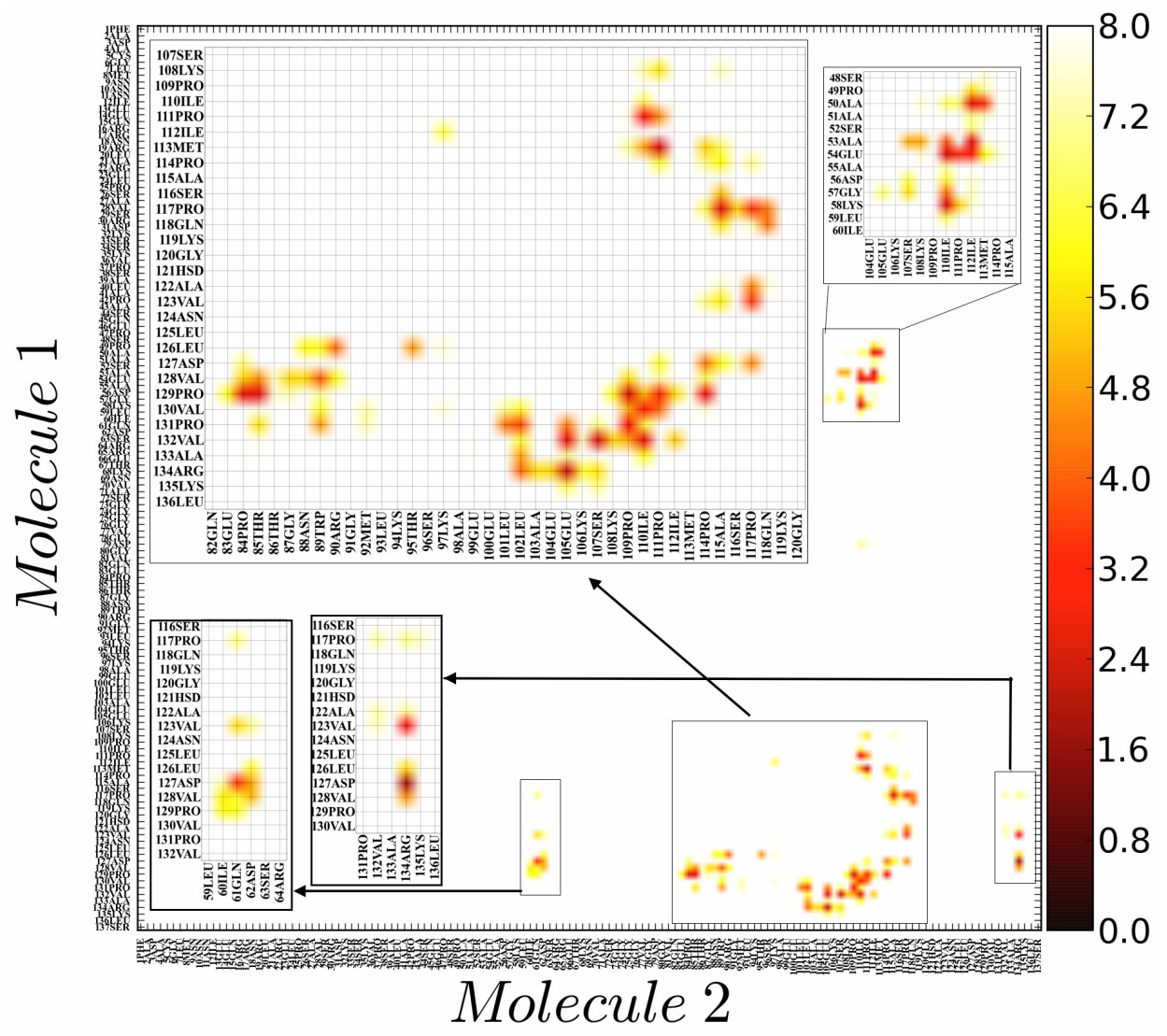


Figure 3.5: Inter-chain distance map of the stable (+Z+Z) B domain dimer in water. The average distance between two chain residues was computed over the last 10 ns of the simulation data. Distances more than 8 Åare not presented.

ARG65, and LYS68 residues involved in interactions between two chains.[†] Overall, charged residues, especially positive ones, were found to be involved in binding pocket for dimer structures in the TMAO solutions, where they interact either with oppositely charged atom from other chains or with TMAO molecules in the pocket to reduce the electrostatic penalty

[†]Note that in native conditions, the B domain linked to the rest of the Drp1 protein and thus its two ends are connected other residues in the Drp1 sequence. Therefore, end residues are not able to be at the centre of dimer pockets in native conditions. Thus, contacts involving PHE1 in the disconnected B-domain (as shown in Figure 3.6 A and D) would not play a role in native Drp1 association.

of having a free charge in the binding pocket.

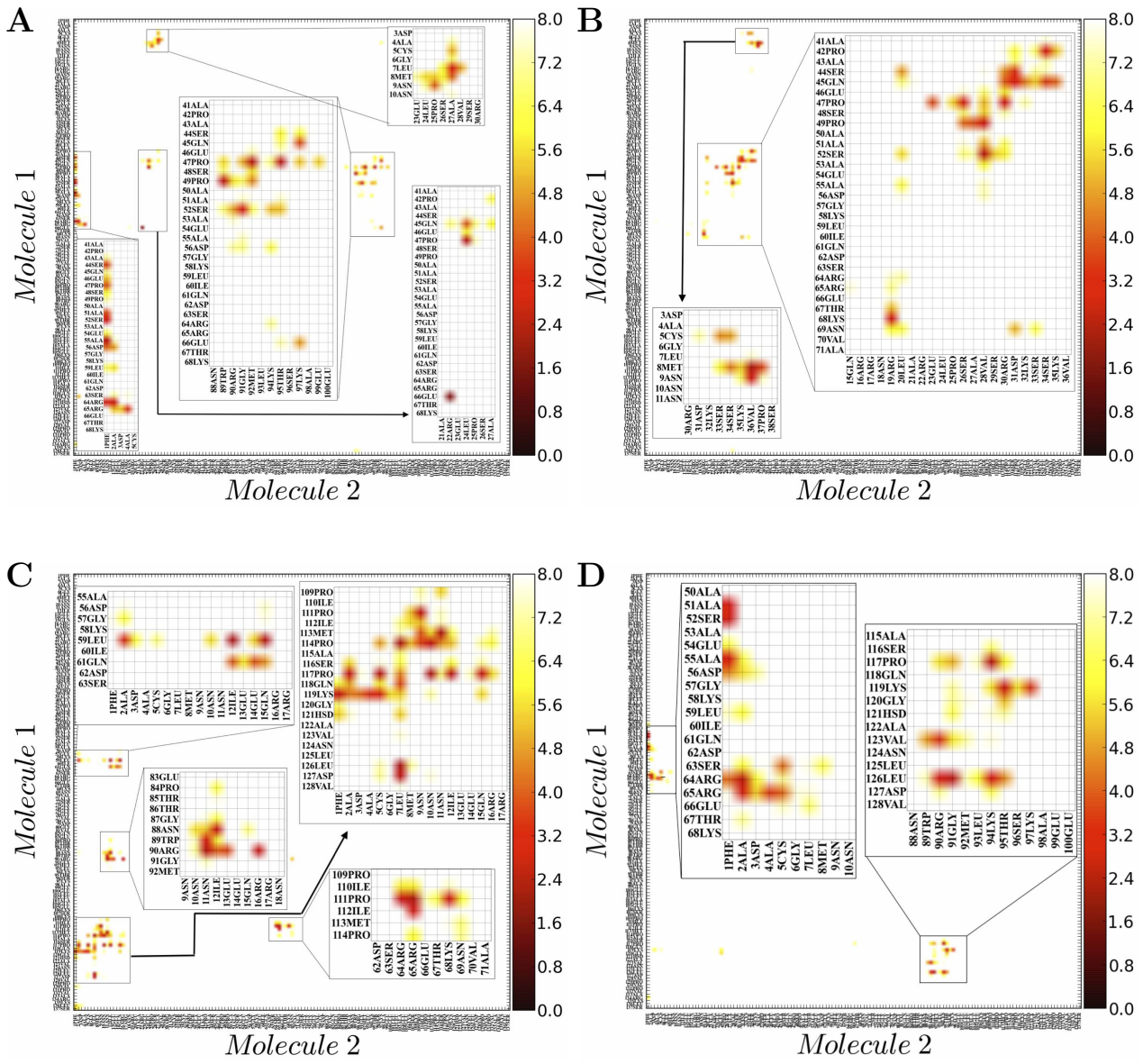


Figure 3.6: Inter-chain distance map of associated B domain pairs for the four most stable dimer structures (A:-Y+Z; B:-Y+Y; C:+Y+X; D:+Z+Z) in the presence of TMAO molecules. The average distance between two chain residues was computed over the last 10 ns of the simulation data. Distances more than 8 Åare not presented.

3.4 Concluding Remarks

Using molecular dynamics simulations, the structure and possible dimer formation of the variable B-domain in Drp1 protein, a member of dynamin superfamily, were investigated. This variable domain was found to adopt a disordered conformation which did not undergo a significant structural transition in the presence of TMAO, a model stabilizer osmolyte, or in response to the formation of homodimers. This is in contrast to the proposed auto-inhibition role for this domain that has been implicated in a similar domain in classic dynamin, another member of this membrane remodeling superfamily. Further analysis indicated that positively charged residues on the B-domain surface play an essential role in its intermolecular association. Moreover, direct interaction with polar TMAO osmolyte was found to reduce the electrostatic repulsion and to mediate and stabilize a population of possible homodimer configurations (four different stable intermolecular associations were identified in the presence of TMAO, whereas only one was found for pure water solution). This suggests that molecular agents which modulate the interactions of Arginine and Lysine residues on the B-domain surface may play a key role in induction of self-oligomerization of the B-domain. As a part of the Drp1 protein, this triggered self-assembly of the B-domain may possibly facilitate oligomerization of Drp1 in physiological conditions. Interestingly, hydrolysed GTP (GTD) from another domain of Drp1 and cardiolipin, a lipid that is unique to the mitochondrial membranes associated with Drp1, both have a negative charge and might act in an analogous way to the TMAO molecules in this simulation study and in the experimental studies of our collaborators in the Hill group. This may also explain the specificity of Drp1 to mitochondrial membranes and relation between hydrolysis activity and protein self-assembly in native conditions. Thus, this work may shed light on the structural and functional assembly of the variable domain in Drp1 protein in its native environment. Future simulation studies of B-domain solutions with cardiolipin and with GTD will further test these ideas.

3.5 Appendices

3.5.1 Forcefield Examination for TMAO molecule

The all-atom model for solvated TMAO molecules in this work was obtained from the Kast model [87] with some optimization done by *Garcia et al* [88] to reproduce the experimentally observed osmotic pressure for TMAO solutions. In comparison to this model, final state for dimer simulation in TMAO solution was continued for another 50 ns using CHARMM General force field (CGenFF2b8) [95]. Results obtained for the two distinct force fields for TMAO were shown in Table 3.5.

CHARMM Model			
Distance	Avg # of Waters	Avg # of TMAO	Water: TMAO
< 4 Å	671 ± 6	36 ± 3	18.6:1
4–8 Å	923 ± 13	101 ± 3	9.1:1
8–12 Å	1187 ± 14	107 ± 1	11.1:1
12–16 Å	1465 ± 20	131 ± 3	11.2:1

Garcia Model			
Distance	Avg # of Waters	Avg # of TMAO	Water: TMAO
< 4 Å	706 ± 6	51 ± 1	13.9:1
4–8 Å	988 ± 6	97 ± 1	10.2:1
8–12 Å	1247 ± 15	113 ± 1	11.0:1
12–16 Å	1537 ± 26	140 ± 4	11.0:1

Table 3.5: Solvent distribution in radial shells from the B domain backbone for two TMAO models. Water oxygen and TMAO nitrogen within a certain distance from the protein backbone for B domain in monomer and dimer simulation. The error estimation of the data was obtained by dividing the 50 ns of simulation trajectory into five equal batches and computing the averages and standard deviations. Ideal water:TMAO ratio is 11:1 for dimer solutions.

CHAPTER 4

Disorder to Order transition in IDPs

A portion of this work has been published in a peer-reviewed scientific journal and this chapter was reproduced in part with permission from,
"Benjamin P. Partlow [†]; Mehran Bagheri [†]; James L. Harden; David L. Kaplan; Biomacromolecules 2016, 17, 3570-3579".

DOI: 10.1021/acs.biomac.6b01086

[†]First two authors contributed equally to this work

4.1 Introduction

One of the primary characteristics of functional disorder is induced folding, which has biological consequences for some IDPs/IDRs. This transition is often triggered via protein-protein interactions when the disordered region is positioned next to its binding partner, the so-called coupled-folding scenario (*c.f.* the SNARE protein explained in Section 1.3.2). Alternatively, such a structural transition might be driven by physicochemical conditions of the polypeptide environment, such as solution pH and ionic strength [96] (*c.f.* Appendix A) which generally play an essential role in folding processes of both ordered and disordered proteins. More complex order-disorder transitions in multi-chain systems are also dictated by specific protein-protein interactions that are modulated by the solution environment. One such example occurs in concentrated solutions of silk proteins, which undergo local structural transitions when they are extruded from the silk gland in a spider or a silkworm during a process called *spinning*. In the spinning process, the silk proteins are initially in random coil conformations in concentrated liquid form. Subsequently, while transported along a narrowing tube with gradients in pH and ionic strength, they change conformation into semicrystalline fibers that contain highly oriented intermolecular beta sheet structures. Figure 4.1 demonstrates the biomolecular machinery of silk fiber production for a silkworm and a spider silk gland.

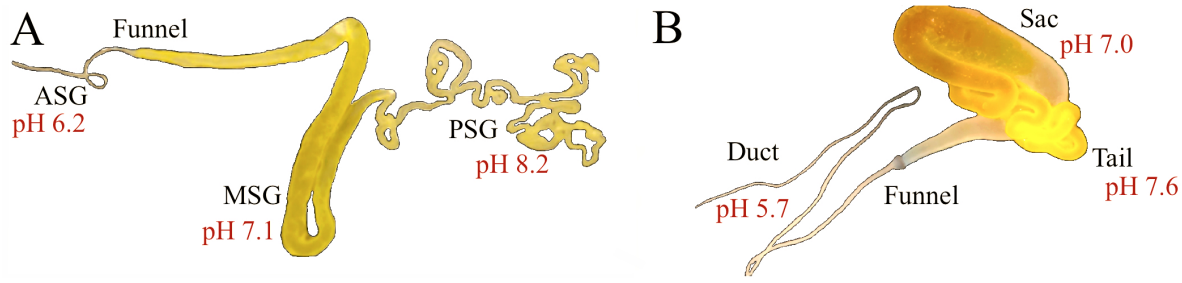


Figure 4.1: Schematic image of a silkworm and a spider silk gland. A) Bombyx mori silk gland indicating the posterior silk gland (PSG), middle silk gland (MSG), Funnel, and the anterior silk gland (ASG). B) Spider silk gland (major ampullate gland) illustrating the Tail, Sac, Funnel, and the Duct. Adapted from [97].

Silk proteins are secreted into the PSG of the bombyx mori silk gland, and the corresponding Tail of the spider gland, and are stored in the MSG and Sac, respectively, at a very high concentration, referred to as *dope*. The dope is drawn through the Funnel where it forms a gel, due in part to the drop in pH and changes in salt concentrations. In addition, the dope is elongated while it is flowing through the ASG/duct, which is narrowing consid-

erably. This elongational flow is also thought to play a role in the process. Ultimately, silk dope is extruded from the gland and exposed to the outside (where it dehydrates), leading to formation of a solid fiber with high beta-sheet content.

Silk fibers have been studied extensively during past century due to their impressive properties such as high strength and extensibility [98, 99], as well as their biocompatibility [100] and biodegradability [101]. Despite a significant amount of studies on the variety of physicochemical elements involved in the spinning process, such as changes in salt concentration, decreases in pH [102], dehydration, and physical forces, the mechanisms that govern the spinning process are still not completely elucidated [97, 103]. Herein, the contribution of the silk protein sequence that determines inter-chain interactions has been investigated using molecular dynamics simulation to elucidate some aspects of the spinning process.

The silk proteins of silkworms, the so-called fibroins, consist of hydrophilic linker domains and highly repetitive motifs which are involved in beta-sheet crystallites. For the Bombyx mori silkworms, the following characteristic repetitive amino acid sequence is found in these motifs,

$$\texttt{GAGAGS GAAS [GAGAGS]}_n \texttt{ GAGAGY GAGVGAGY GAGYGAGAGAGY}$$

where n can vary from 1 to 11 for different motifs [104]. The initial part of the motif, rich in -Gly-Ala-Gly-Ala-Gly-Ser-, has been found to form beta-sheet conformations in silk fibers [105, 106]. On the other hand, a clear role of the tyrosine-rich region (GY...GY) is not well understood, despite several experimental studies [107, 108].

Tyrosine residues are able to form relatively strong non-covalent π-π interactions via their aromatic rings and act as structural stabilizers, as observed in many DNA and protein structures, such as in the DNA double helix [109] and in beta-hairpins [110]. They often occur in clusters. For example, more than half of unique proteins in Protein Data Bank [20] were reported to have colocation of tyrosines with more than two interacting aromatic groups in their crystallized structure [111].

In a recent study carried out by our collaborators at Tufts University (See e.g. Partlow et al [112]), a strong correlation between beta-sheet formation and the local environment of the tyrosine residues was reported. They found that during a structural transition of a regenerated silk solution from random coil conformation to beta-sheet structure (as determined by FTIR), the peak corresponding to the maximum tyrosine fluorescence emission spectra varied accordingly (See Figure 4.2), suggesting a change in local environment surrounding

tyrosine residues, presumably due to a change in solvation of aromatic rings as a result of colocalization of tyrosines and their subsequent involvement in beta-structure formation.

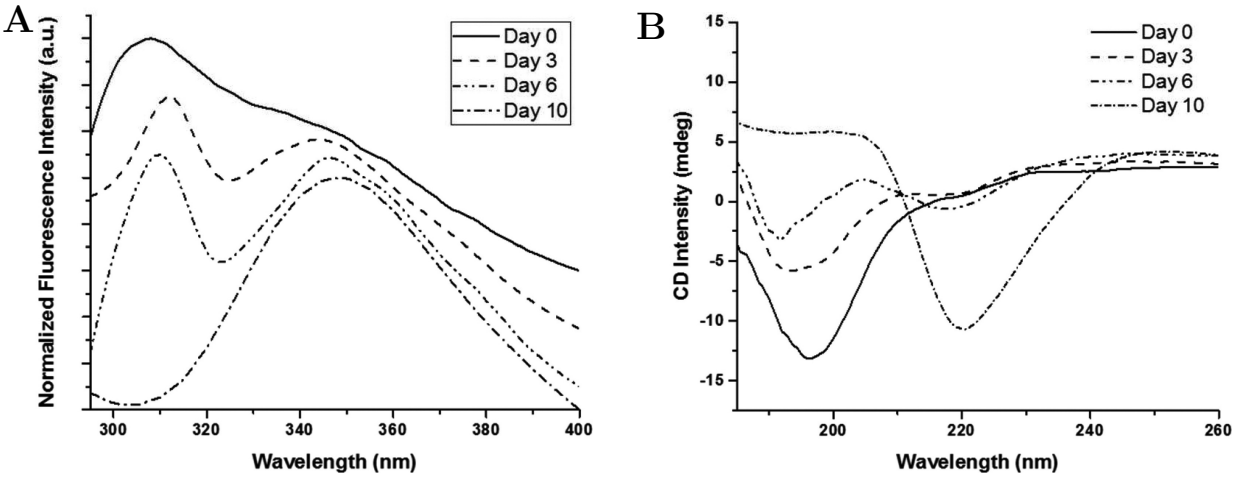


Figure 4.2: Fluorescence and structure of regenerated silk solution over time. (A) Intrinsic fluorescence of the silk solution over time indicates that tyrosine residues, with an emission of ~ 305 nm are dominant immediately after dissolution lyophilized silk fibroin. However, during the course of self-assembly, the tyrosine emission disappears and the tryptophan fluorescence with emission at ~ 345 nm dominates. (B) CD spectra reveal a predominantly random coil structure, with a characteristic minima near 195 nm, in fresh solution that slowly transitions to a gel rich in a beta-sheet structure, as indicated by depression around 218 nm. Figures and captions are reproduced with permission from [112].

Moreover, Partlow et al [112] showed that the beta-sheet formation can be accelerated from few days to under an hour by covalently cross-linking tyrosine residues in regenerated silk solutions (See Figure 4.3 A). This modification transformed transient tyrosine association into permanent colocalization and also induced regenerated silk solutions to crosslink into hydrogels via the formation of intramolecular dityrosine crosslinks. The formation of dityrosine was independently confirmed and quantified by measuring fluorescence intensity at 315 and 415 nm corresponding to a dityrosine excitation and emission signal, respectively. The resulting hydrogels were found to have elastic properties that depended on the cross-link density and solution molecular weight [113]. Interestingly, these silk hydrogels were found to form significant beta-sheet content rapidly upon dehydration, without necessary extra treatments typically required to induce beta-sheet structures in the non-crosslinked silk solutions. Moreover, higher beta-content was induced by the dehydration process for hydrogels with higher initial cross-link density. Figure 4.3 B shows beta-content as a func-

tion of cross-link density as assessed by the intensity of dityrosine fluorescence. Although no inherent dityrosine could be detected in Bombyx mori fibroin, evidence of dityrosine bonds has been reported for other worm silks, such as Tussah silk fibroin and marine silks from caddisfly larvae [114, 115], as well as in web silk proteins of several species of spiders [116]. This anecdotal evidence suggests that the tyrosine residues may have an essential role in the beta-sheet formation of the fibroin molecules in silk spinning process.

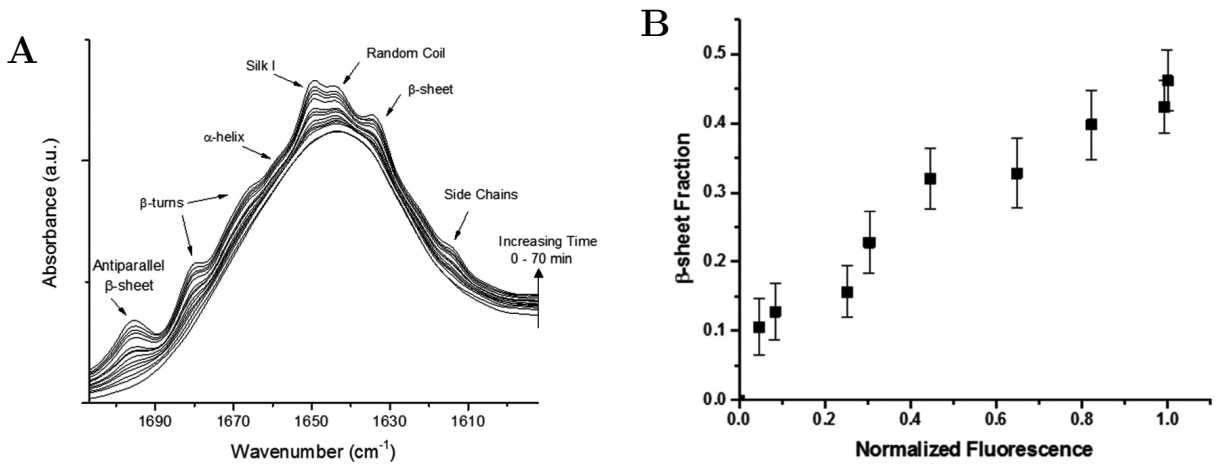


Figure 4.3: Structure of cross-linked silk hydrogel as function of time and cross-link density. (A) Time-dependent FTIR spectra indicates an increase in bands associated with beta-sheet and crystalline silk I conformations immediately following tyrosine cross-linking and up to 70 min after. (B) Relationship between the relative fluorescence and beta-sheet content clearly show that maximizing cross-linking results in higher crystallinity. Figures and captions are reproduced with permission from [112].

In addition to the possible contributions of tyrosine residues, silkworms and spiders experience hydrodynamic stress during flow in the gland tube that can trigger the structural transition from random coil silk fibroins in the spinning dope into highly oriented fibers [117]. One model has proposed that fibroin molecules preferentially form micelle-like structures with a hydrophobic crystalline core which is protected from the aqueous environment by the hydrophilic linker regions and N- and C- termini [118]. In the spinning process, it has been suggested that these micelles aggregate into larger globule-states which are elongated, due to the flow, to form fibrils when finally extruded [119]. To what extend this flow elongation is important and how it affects beta-sheet content and the arrangement of beta-strands is still not clear.

The present study is an attempt to investigate the possible roles of both chain elongation

and interactions between tyrosine residues in the templating of beta-sheet structure in silk fibroin solutions. To achieve this goal, a series of simulations were performed using different sets of silk-like peptides which were designed based on repeat motifs from B. mori silk fibroin. First, tyrosine cluster formation in the silk-mimetic peptide solutions was examined using two different concentrations. The effects of transient and permanent tyrosine associations on the stability of beta-sheet structure were also studied. In addition, a set silk-like peptides elongated by various tensions were simulated as a model for possible effects of elongational flow on beta-sheet formation during the silk spinning process.

4.2 Simulation and Analysis Methods

4.2.1 General simulation details

The Influences of tyrosine residues on silk protein interactions were investigated at the molecular level through simulations of two sets of short silk-mimetic peptide chains. In one set, tyrosine participation in self-assembly of transient inter-chain contacts was investigated, and in the other set, their role in stabilizing beta-sheet structure was examined. Furthermore, a separate series of simulations were carried out using silk-mimetic peptide prearranged in beta-sheet structure (parallel and antiparallel) under applied tensions ranging from 0.1 kT/Å to 12 kT/Å to investigate the role of elongation stress in the spinning process. All sets of peptide sequences were designed based on the highly repeated motifs in silk, [GAGAGS] and [GAGY], as well as control sequences in which tyrosines were replaced by serine residues. Starting conformations were constructed using a custom designed Python code developed by the author (See Appendix B).

In all simulations, peptide chains were solvated within a water box with periodic boundary conditions using the VMD package [52]. The simulation box was extended to maintain a minimum of 15 Å between the peptide atoms and the boundary of the simulation box at all times. All Molecular Dynamic (MD) simulations employed a Langevin thermostat and barostat and were conducted in pure water under an isothermal-isobaric ensemble (NPT) at 300 K and 1 bar. The CHARMM36 force field [68] with the protein backbone improvement [69] and optimized dihedral angles [86] was adopted for the peptides and the TIP3P model was used for the water molecules [70]. A cut-off distance of 12 Å was used for electrostatic and Lennard-Jones interactions and electrostatic interactions were computed via

the particle-mesh Ewald (PME) algorithm. All atomistic simulations were carried out using NAMD 2.10 [71] with a time step of 2 fs.

4.2.2 Equilibration of silk solution and tyrosine self-assembly

Interchain tyrosines contacts were studied for silk solutions using following silk-mimetic peptide sequences:

```
Silk-Y: Y GAGAGSGAGAG Y GAGAGSGAGAG Y
Silk-S: S GAGAGSGAGAG S GAGAGSGAGAG S
```

where in the second, control sequence, the tyrosine residues are replaced by serine residues. Solutions with different concentrations of the Silk-Y and the Silk-S variants were simulated to characterize aromatic and peptide-peptide interactions by comparing them as a function of concentration and sequence. To initialize multi-chain simulations, configuration ensembles for silk-mimetic peptides were obtained from single chain simulations. First, individual peptides (Silk-Y and Silk-S) were constructed in a fully extended configuration and separately simulated for 100 ns as single chains. Then, distinct configurations from the last 50 ns of the simulations were selected and used as initial chain conformations for subsequent multi-chain simulations. Three sets of simulations were conducted, each for 200 ns, for each peptide variant at two concentrations of 3% and 9% wt.

In the multi-chain simulations, an inter-chain association event was noted between particular pairs of residues when their spatial separation was less than a chosen cutoff distance. In particular, a tyrosine-tyrosine association event was recorded when the distance between centers of the aromatic ring is less than 6 Å, while a serine-serine contact can have, at most, a distance of 5 Å between the sidechain centers of mass. In this analysis, tyrosine-tyrosine and serine-serine associations were assessed in terms of frequencies and lifetimes of these association events. Also, short-lived associations which have lifetimes shorter than 0.1 ns were ignored and sampling was carried out for the last 50 ns of individual simulations.

4.2.3 Stability of silk peptide in beta-sheet structure

To evaluate the role of the tyrosine residues on the stability of beta-sheet structure, two complementary peptide sequences were employed that were initially prearranged in an antiparallel beta-strands configuration. During the simulations, the beta-sheet content of these

structures was monitored, especially for the central region where two tyrosine residues were located. These complementary sequences were as follows:

```
Silk-mY1: GAGAGSGAG Y GAG Y GAGAGSGAG
Silk-mY2: GAGAGSG Y GAG Y GAGAGAGSGAG
```

These peptide sequences are complementary in that they allow the tyrosines from one strand to be adjacent to those in the other strand when prepared in an antiparallel beta-sheet configuration. Two such antiparallel beta-sheet strands were employed to construct a four-chain beta-sheet bundle conformation by positioning each strand so that the glycine residues were facing each other. A schematic of this configuration is illustrated in Figure 4.5 A. Two other versions of this four-chain beta-sheet bundle were prepared for a better assessment of tyrosine effects. In one version, nearest-neighbour tyrosine residues were permanently connected to mimic a di-tyrosine cross-link and in the other one, serines replaced all tyrosine residues as control peptides. Six sets of simulations for each of these three types of peptides were performed for 25 ns each. During these simulations, a small tension of 0.1 kT/Å was applied to the C_α of the terminal residues of individual chains in order to avoid large end fluctuations in these short peptides and to mimic real chain behaviour in which this repetitive motif is embedded in a much longer protein sequence.

In the last set of studies, the Silk-Y peptide introduced in the previous section (Section 4.2.2) was pre-arranged into three-stranded beta-sheet structures (both parallel and anti-parallel variants), and two such three-chain beta-sheet strands were employed to construct a six-chain beta-sheet bundle conformation by positioning each strand so that the glycine residues were facing each other. Both parallel and anti-parallel assemblies were simulated under five different tensions (0.1 kT/Å, 1 kT/Å, 3 kT/Å, 6 kT/Å, and 12 kT/Å) to investigate how extensional forces experienced by the silk dope during fiber spinning influence the beta structures. Two sets of simulations, 25 ns each, were carried out for each beta-sheet structure variant during application of the desired tension to the alpha-carbon of the terminal residues. The time-dependent root mean squared displacements (RMSD) of the peptide backbone atoms relative to those in an aligned, ideal beta-sheet bundle (the initial condition) were found to converge to steady-state conditions within 10 ns of simulation time in all cases. The beta-sheet content was measured over the last 15 ns of the simulations (the sampling period) using the STRIDE algorithm [72] implemented in VMD software [52]. This algorithm assigns beta structure based on beta-compatible dihedral angles and hydrogen

bonds. In particular, beta content was assessed in both "Extended" and "Isolated-bridge" states, where the former occurs when there are two or more consecutive inter-strand hydrogen bonds and the latter refers to events where single inter-chain hydrogen bond are observed.

4.3 Results and Discussions

4.3.1 Tyrosine colocation in solution

Several experimental studies suggest a correlation between tyrosine self-assembly and secondary structure of reconstituted fibroin chains [120, 121]. As indicated in Figure 4.2, an increase in the content of beta structure in regenerated silk coincides with a noticeable change in the local environment of tyrosine residues, suggesting tyrosine involvement in the formation of higher order structures that result in colocalization of the tyrosine residues. Particularly when tyrosines are covalently coupled via dityrosine bonding, beta-sheet secondary structure starts to form almost simultaneously upon dehydration of the fibroin solution (See Figure 4.3). Here, molecular dynamics (MD) simulation was employed to probe the self-assembly process at the atomistic level and to evaluate the extent to which tyrosine residues can stabilize beta sheet structure in silk-mimetic peptide solutions.

Solutions of silk-like peptides, including control peptides with serine substitutions for tyrosines, were simulated at two different peptide concentrations to examine the role of the tyrosine amino acids in transient inter-chain associations. These associations were reported as the ratio of selected residues that were within a cutoff separation (6 Å for tyrosines or 5 Å for serines in the control sequence) to the total number of the residues. Results for the silk-like peptide containing tyrosines and control ones with replacement serines are shown in Table 4.1. At low concentration (3% wt.), there were slightly more associated residues in the presence of tyrosines; the percentage of association increased from 8 $\pm$ 2% for the peptides with serine substitutions to 22 $\pm$ 10% for the peptides with tyrosines. Moreover, longer average association lifetime was observed for tyrosine pairs (2.1 $\pm$ 1.2 ns) compared with replaced serine pairs (0.5 $\pm$ 0.3 ns). Tyrosine pairs also showed a noticeable increase in the total number of long-lived associations (defined as those which have lifetimes longer than 0.5 ns, the average apparent lifetime for a serine pair association). While there was only 1 $\pm$ 1 serine pair that was associated for more than 0.5 ns in the control sequence, there were 8 $\pm$ 1 tyrosine pairs associated for more than 0.5 ns in the target sequence.

	Low Concentration		High Concentration	
	Silk-Y	Silk-S	Silk-Y	Silk-S
Percentage of residues associated	22 ± 10%	8 ± 2%	44 ± 7%	12 ± 4%
Total association time per selected residue	2.1 ± 1.2 ns	0.5 ± 0.3 ns	2.3 ± 0.3 ns	0.8 ± 0.3 ns
Number of association longer than 0.5 ns	7 ± 6	1 ± 1	43 ± 11	12 ± 4

Table 4.1: Silk-like peptide association at low and high concentration. Statistical results given for two silk-like peptides, Silk-Y and Silk-S. The Silk-Y variant has three tyrosine residues, one in the middle and one at each end; these tyrosines were replaced by serines in the Silk-S variant. Results were provided for low and high concentration solutions (3% and 9% wt). For each configuration, three independent simulations were executed, where the average and standard deviation were calculated over last 50 ns. Table results are adopted with permission from [112].

As the silk-like peptide concentration was increased, the effects of tyrosine were significantly enhanced. Simulation results conducted at 9% wt indicate a notable increase in the percentage of association for tyrosine pairs compared to replaced serine (44 ± 7% to 12 ± 4%). Also, the number of long-lived associations (lifetime >0.5 ns) was considerably increased; there were 43 ± 11 corresponding tyrosine pairs in contrast to 12 ± 4 for the serine pairs in the control peptide. On the other hand, no significant difference was found between low and high concentration for the average lifetime of tyrosines association (2.1 ± 1.2 ns at 3% wt and 2.3 ± 0.3 ns at 9% wt silk-like peptide solution); this indicates that the effect of increasing concentration (at least within this range) on the translational contribution to entropy of these associations can be ignored. Therefore, the association free energy (and consequently their stability of tyrosine pairs), is essentially independent of tyrosine concentration. However, associated pairs of tyrosine did show substantially longer lifetime than serine pairs (0.8 ± 0.3 ns) at higher concentration, implying a concentration dependence to association kinetics.

Tyrosine residues are able to interact via their aromatic ring in multiple ways, such as through the well-known π-stacking and T-stacking [122], as well as through hydrogen bonding via the OH group in their sidechains. These various interactions contribute to the association events observed in the simulations of tyrosine-containing silk-like peptides. Distinct instances of these tyrosine interactions were captured in snapshots from multi-chain simulations for peptides incorporating tyrosine residues and shown in Figure 4.4 B. Importantly,

these interactions facilitate tyrosine association between different chains, thereby mediating inter-chain association that acts to position chains adjacent to each other (See Figure 4.4 A).

This work on short silk-mimetic peptide suggests that tyrosine residues are partially responsible for the self-assembly of the chains in the native silk dope by aromatic and other sidechain interactions. The observed transient inter-chain tyrosine association reduces relative chain fluctuations and may thereby enhance intra-chain hydrogen bond and subsequent beta sheet formation in native silk protein solutions during dehydration. This is consistent with an experimental study that showed aromatic residues are able to increase stability beta-structure of protein variants containing glycine [123]. This is further investigated in next section.

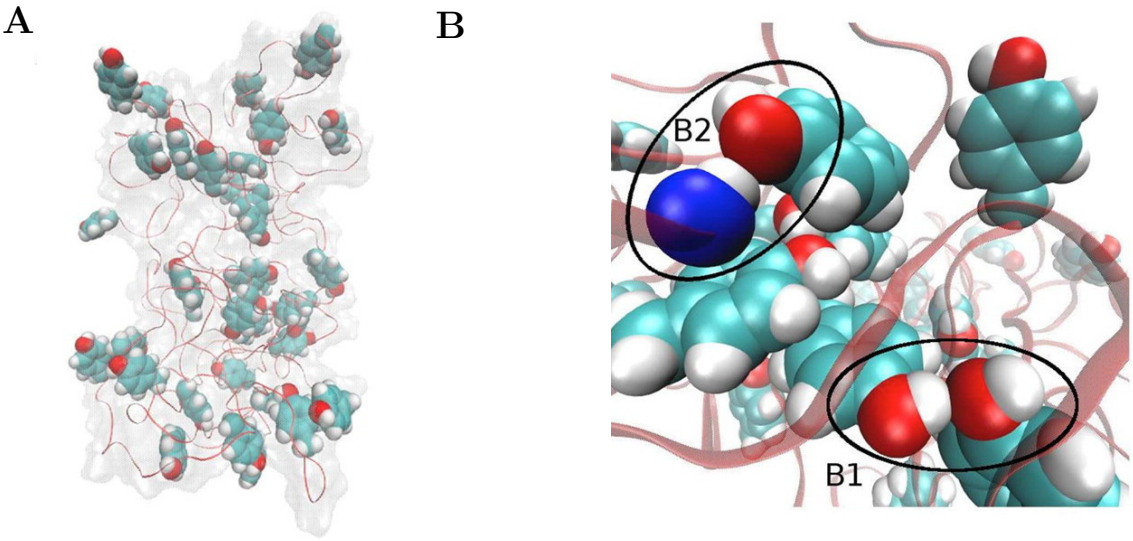


Figure 4.4: Representative transient network structure of a 9 wt % solution of silk-mimetic peptide containing tyrosine. Following simulation for 200 ns, (A) the tyrosine residues in the silk-mimetic peptide chains tend to self-assemble into metastable clusters. (B) In addition to tyrosine π-π interactions, these residues are able to interact with each other (B1) or with the chain backbone (B2) via their OH groups. Synergistic effects of multiple types of interactions could further enhance the stability of these associations. Figures and captions are reproduced with permission from [112].

4.3.2 Tyrosine residues promote beta-sheet structure

The role of tyrosine residues in beta-structure formation is still a matter of controversy, although several experimental studies suggest that tyrosine residues promote and stabilize

beta-sheet secondary structure [123, 124]. To determine the effect of tyrosine residue on extended beta-sheet structure stability, multiple simulations were conducted using silk-like peptides (the silk-mS series) which were initially pre-formed in four-chain beta-sheet bundles incorporating of two pairs of antiparallel beta-strands positioned with glycine residues facing each other. A schematic of this configuration is shown in Figure 4.5 A. Three different versions of this configuration under weak tension (0.1 kT/Å) were studied: a) configurations where tyrosine residues were covalently bonded, b) configurations where tyrosine residues were free to interact with neighbouring residues, and c) configurations where tyrosines were replaced by serines. In each case, six trails of such assemblies were equilibrated for 25 ns and the stability of the pre-formed beta structures were monitored. Snapshots of the configurations for equilibrated bundles with cross-linked and free tyrosines are shown in Figure 4.5 B and C, respectively.

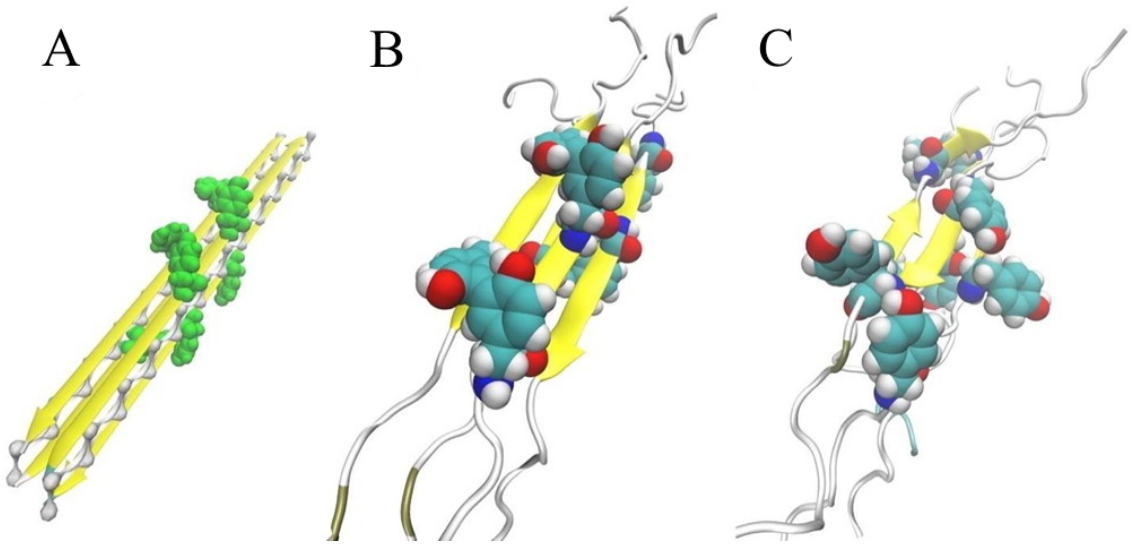


Figure 4.5: Representative conformations of silk-mimetic peptide bundles. Initial beta-conformation and conformations taken near the end of simulation of four-chain silk-mimetic peptide bundles. (A) Initial conformation of an ordered beta-sheet bundle composed of two antiparallel beta strands (yellow arrows) facing each other from the glycine side (white), in which tyrosine residues (green) are adjacent to each other. (B) Post-simulation conformation of a bundle with cross-linked tyrosine residues, showing stabilized beta-sheet structure (yellow arrows) in the central portion of the bundle. (C) Post-simulation conformation of a bundle with free tyrosine residues, showing relatively weaker, fluctuating beta-sheet structure (yellow arrows) in the central portion of the bundle. Figures and captions are reproduced with permission from [112].

The stability of beta-sheet bundles were examined by measuring average beta content for

the last 15 ns of simulation when backbone atom root mean square displacement (RMSD) was in steady-state conditions. Results of average beta content are given in Table 4.2. Assemblies that incorporated cross-linked tyrosine showed significantly higher beta content (37 ± 8%) than either the free tyrosine (20 ± 4%) or serine substituted (17 ± 4%) configurations. Moreover, in the core region where tyrosine residues were concentrated, beta content was substantially elevated for the cross-linked tyrosine (72 ± 16%) and free tyrosine (35 ± 12%) configurations; whereas in the serine control configurations, there was no significant difference between core and full chain averages (17 ± 4% to 22 ± 3%). Therefore, we may conclude that the presence of central tyrosines stabilized beta structure in these chain assemblies in comparison to the serine substituted peptide control assemblies. This is more clear in Figure 4.6, which shows a beta-strand heat-map indicating the propensity of individual amino acids to maintain stable beta-sheet conformations.

	Cross-Linked Tyrosine	Free Tyrosine	Serine Substitution
Chain average	37 ± 8%	20 ± 4%	17 ± 4%
Core average	72 ± 16%	35 ± 12%	22 ± 3%

Table 4.2: Average beta content is given for all residues (chain average) and for the central five residues (core average) as determined from simulation data during last 15 ns of simulation for six independent trails of configurations with cross linked tyrosine, free tyrosine, and with serines substituted for tyrosines. Table results are adopted with permission from [112].

Covalently linked tyrosines were able to enhance hydrogen bonding between strands and stabilize the beta-sheet structure by reducing lateral chain fluctuations (Figure 4.6 A). This is in agreement with the improvement of beta-sheet formation via cross-linking tyrosine residues that Partlow et al observed in their regenerated silk experiments [112]. In addition, free (noncross-linked) tyrosines (Figure 4.6 B) also were able to increase beta structure stability, in comparison to the control peptide with substituted serine peptide (Figure 4.6 C). Tyrosines enhance beta structure both by shielding the backbone from solvent and by reducing chain fluctuations during their transient association. This is consistent with the stabilizing role of aromatic sidechain occuring when they are paired with cross-strand destabilizer residues (glycine) in beta-sheet structures [123]. Overall, we may conclude from these simulations that tyrosine association improves hydrogen bonding between strands and stabilizes beta-sheet content in the core region in comparison to the serine control.

As a result, it is suggested that tyrosine cross-strand association, although transient

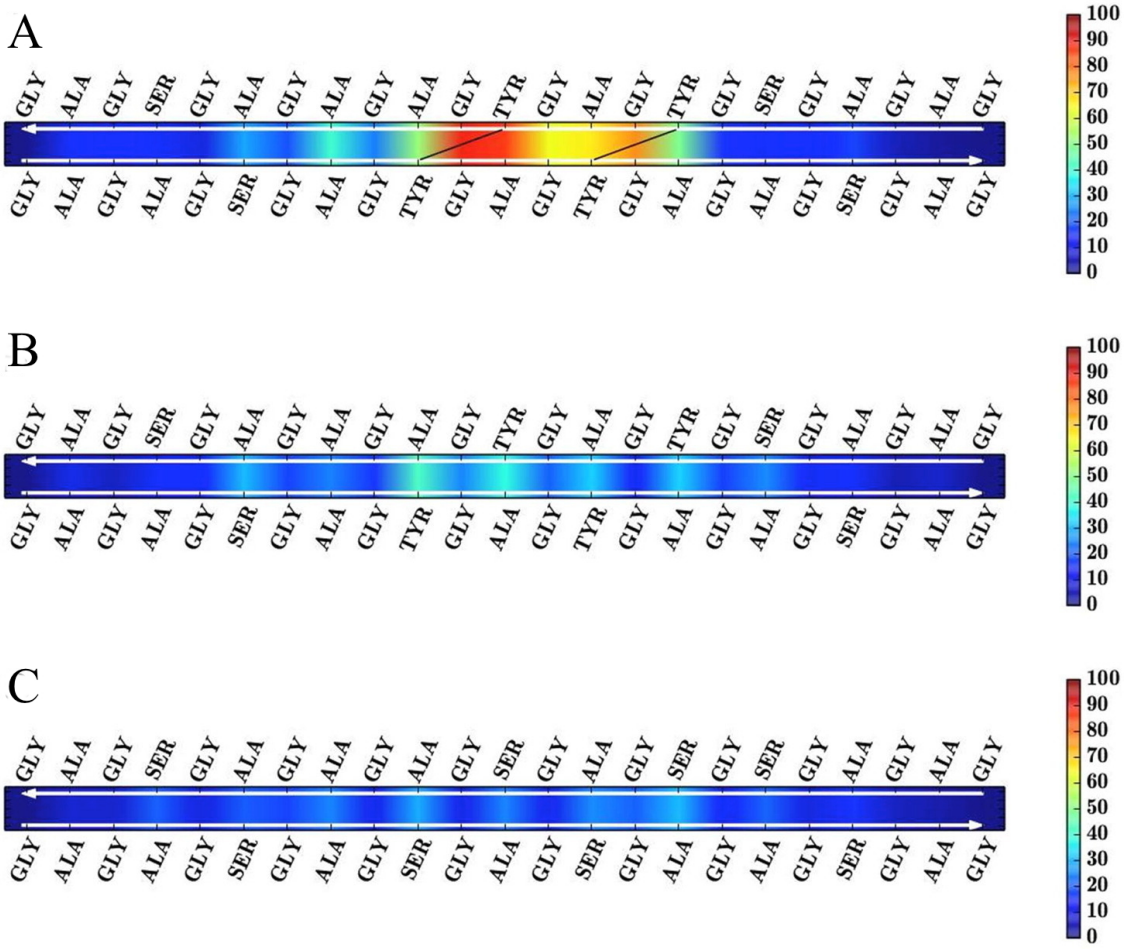


Figure 4.6: Average propensity of individual residues for beta-sheet conformation over six trials. Beta content for individual residues averaged over two beta-sheets in the bundle during the last 15 ns of simulation for six independent trials and presented as a color map. (A) cross-linked tyrosine residues form beta structure for residues in the core region which are highly stable (shown in dark red) and promote metastable beta content for the rest of the chain. (B) Non-cross-linked tyrosine residues fail to maintain initial beta structure but preserve transient beta structure in the core region near the tyrosine residues. (C) When tyrosine is replaced with serine, the initial beta structure is disrupted completely and does not show any tendency for beta structure to reform. Figures and captions are reproduced with permission from [112].

and not covalently fixed, are still able to facilitate transient beta-structure in their vicinity. Although the effect of tyrosine in the free state was not as significant as the cross-linked state for stabilizing the beta structure, these stabilizing effects are relevant considering the large fluctuations inherent with a short peptide. Moreover, this stabilization may be substantially

enhanced by other factors involved in the native spinning process, such as flow-induced chain elongation and dehydration. These results are consistent with other experimental work on silk fibroin proteins suggesting tyrosine sidechains do not particularly inhibit crystal formation [125] and that cross-strand pairing with tyrosine residues can rescue beta structure from destabilization [123, 126].

4.3.3 Effects of chain tension on beta-sheet structures

Silk protein is under considerable tension during the spinning process. The silk dope is transported through a narrowing tube that generates shear and elongational stresses on the highly concentrated silk liquid, thereby forming fibers. These fibers are extracted from the end of the tube by the motion of the silkworm head. Previous studies showed that such flows can induce beta-sheet structure in silk solutions [117, 127]. Intuitively, elongating the silk protein during extrusion would align the silk chains and reduce lateral chain fluctuations, consequently inducing beta-sheet structure aligned with the fiber axis [128]. Although the main structural elements in silk fibers are beta-sheet assemblies, it is still unclear if the beta content is predominantly composed of parallel or anti-parallel beta strands. [129]. Here, simulations of chains under tension for pre-formed beta-sheet structures were carried out to investigate how elongational flow might influence silk crystal domains.

Simulations were conducted on silk-like peptides having three tyrosines, two at each end and one in the middle (Silk-Y), that were initially arranged into six-chain beta-sheet bundles formed by stacking extended parallel and antiparallel beta sheets of three strands, similar to the four-chain bundle structure shown Figure 4.5 A. These assemblies then were simulated under various tensions ranging from 0.1 kT/Å to 12 kT/Å. For both parallel and anti-parallel beta strands assemblies, two trails of such assemblies were equilibrated for 25 ns and the evolution of beta-structure was monitored. The average beta content over all residues during last 15 ns of the simulation is given in Table 4.3. For both strand orientations, average beta content was found to increasing function of applied tension from 0.1 kT/Å to 12 kT/Å, due to the progressive suppression of lateral chain fluctuations and the concomitant stabilization of inter-strand hydrogen bonds. For the case of pre-formed parallel beta-sheet configurations, the increase in average beta content (from 12% to 46%) was gradual and moderate (but significant). However, for pre-formed anti-parallel beta-sheet configurations, this increase was more dramatic, jumping from 21% to 73% at a tension of 3 kT/Åand reaching nearly full

beta content (92%) at 12 kT/Å tension. This response to applied tension and its dependence on beta strand orientation can be explained by examination of backbone dihedral angles and inter-strand hydrogen bonds under applied tension.

	0.1 kT/Å	1 kT/Å	3 kT/Å	6 kT/Å	12 kT/Å
Parallel strands	12 ± 3%	20 ± 1%	25 ± 1%	38 ± 2%	46 ± 2%
Antiparallel strands	15 ± 1%	21 ± 3%	73 ± 7%	82 ± 2%	92 ± 1%

Table 4.3: Average beta content for silk-mimetic initially in six-chain beta sheet bundles structure under various tension. Average beta content is given for all residues as determined from simulation data during last 15 ns of simulation for two independent trails of beta struture with parallel and antiparallel orientation. 1 kT/Å is about 41 pN at the simulation temparature, 300K. Table results are adopted with permission from [112].

In the absence of any external forces, typical values of dihedral angles (ϕ,ψ) in the most stable state are about (-120°,115°) and (-140°,135°) for parallel and antiparallel beta structure, respectively. Application of tension beyond 1 kT/Å was able to influence the dihedral angles and shift their constrained equilibrium values toward (-160°,160°), close to values corresponding to a fully extended chain conformation (-180°, 180°). The dihedral angle distribution for selected tension values is shown in Figure 4.7. Shifting ϕ to higher negative value and ψ to higher positive value induces a clockwise twist in the peptide bond (in the direction of the strand, from N terminal to C terminal), which affects the stability of inter-strand hydrogen bonds between carbonyls and amines in an orientation dependent manner.

Even in the absence of applied forces, antiparallel beta-sheets have higher stability compared to parallel ones, due to the planarity of their inter-strand hydrogen bonds (See Figure 4.8 A, B). Under applied tension, therefore, the deviation of the hydrogen bond donor and acceptor from their preferential arrangement is more disruptive in parallel beta-sheet configurations than in antiparallel configurations. Although the twist induced by tension deforms cross-strand carbonyls-amine pairs from their preferential planar conformation in antiparallel configurations, these carbonyls and amines are maintained in close proximity as they are driven in the same direction (See Figure 4.8 C). In contrast, these carbonyls and amines move in opposite directions in parallel beta configurations, progressively disrupting inter-strand hydrogen bonds in response to tension (Figure 4.8 D).

Applied tension in our simulations mimics the extensional stress under flow that facilitates orientation of the chain during the spinning process. Interestingly, tension on the polypep-

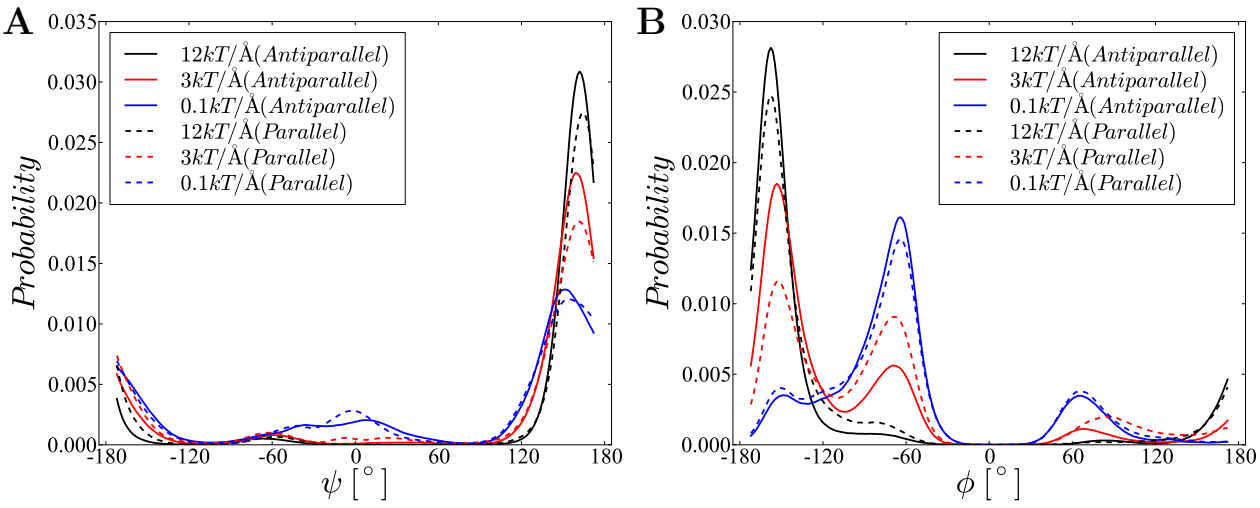


Figure 4.7: Distribution of dihedral angles (A: ψ and B: ϕ) for silk-mimetic protein in two different beta-sheet orientations (solid line: antiparallel; dashed line: parallel) under 0.1 kT/Å, 3 kT/Å, and 12 kT/Å tension. Statistical data was extracted from conformations during the last 15 ns of the simulations.

tide chain does not affect parallel and antiparallel beta-strands equally; rather, antiparallel beta-sheet structures were found to be more stable than parallel ones. This finding is not limited to silk proteins; it should be a generic feature for all beta-sheet forming polypeptides regardless of their amino acid sequence. Antiparallel beta structures are more common than parallel beta-sheet in nature, possibly due to their planar inter-strand hydrogen bond arrangements. However, tension during biological processes may also promote antiparallel beta-sheet structures. To the best of my knowledge, this is the first time that distinct effects of external tensions are reported for parallel and antiparallel beta-sheet structures in protein or peptide assemblies.

4.4 Concluding Remarks

The roles of tyrosine residues and flow-induced extension on the stability of beta-sheet secondary structure in self-assembly of tyrosine-containing silk-like peptides were investigated using molecular dynamic simulations. Tyrosine residues showed high propensity for co-localization in silk peptide solution due to their transient aromatic interactions. As a result, inter-chain associations were significantly enhanced, providing appropriate conditions

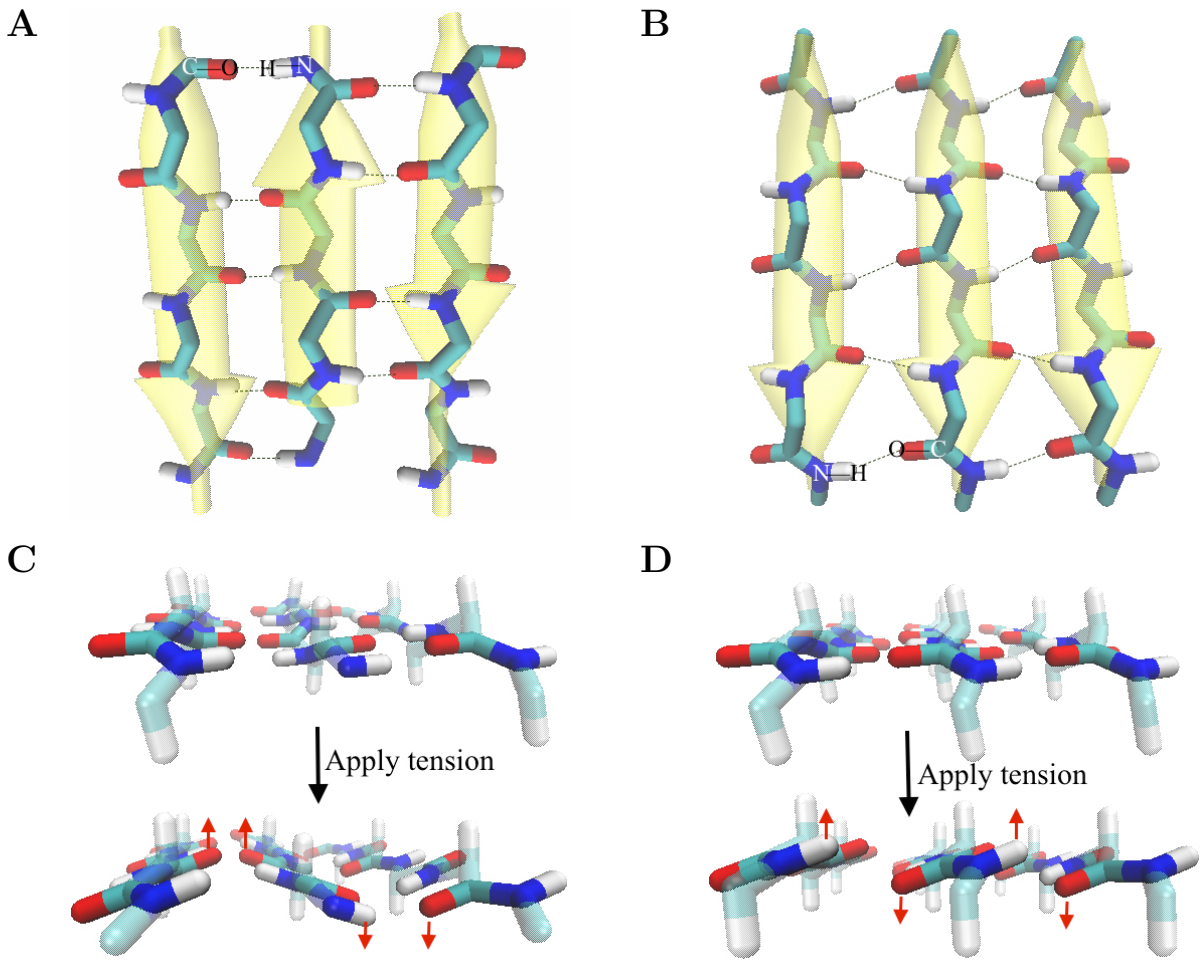


Figure 4.8: Beta-sheet structure details and effects of tension on the polypeptide backbone. The configuration of antiparallel (A) and parallel (B) beta-sheets indicating inter-strand hydrogen bond patterns. The directions of strands are indicated by yellow arrows and hydrogen bonds between backbone carbonyls and amines atom are represented by dashed lines (Oxygen, red; Nitrogen blue; Hydrogen, white). C) The orientation of the backbone carbonyl and amine groups for an antiparallel beta-sheet with backbone dihedral angles of (-130°,130°) and (-160°,160°), obtained in the absence and presence of external tension, respectively. D) Details of backbone orientation for a parallel beta-sheet with similar dihedral angles to part C. Tension induces torsion, clockwise in the direction of the strand (N to C terminus), and moves carbonyl and amine groups which are involved in hydrogen bonds (red arrows). The cross-strand donor-acceptor elements of the pairs move in the same direction in antiparallel beta-sheet configurations, whereas they are driven in opposite directions for the parallel beta-sheet configurations.

for templating beta-sheet formation. In addition, tyrosines in pre-formed beta-sheet bundles were able to stabilize secondary structure locally. Such impact was significantly pronounced upon cross-linking these tyrosine residues. This is in excellent agreement with experimental results where cross-linked tyrosine promotes beta structure efficiently.

On the other hand, applying tension on the parallel and antiparallel beta-sheet assemblies indicates that high tension (>1 kT/Å) influences these structures in a quantitatively different manner, wherein there is significantly higher stability provided in antiparallel assemblies compared to parallel ones, due to the differences in intrinsic orientation of hydrogen bonds in these two beta conformations. Therefore, we expect that extensional forces in the spinning process, due to the extrusion silk protein solution through the narrowing funnel and duct regions of silkworms and spiders, is biased in favour of creating antiparallel beta structure in the resulting fibrils.

This study provides substantive support for a key role of tyrosines and tension along the chain in the beta-sheet formation of silk fibroins through intermolecular associations and manipulation of backbone dihedral angles, respectively. As such, this work has provided insights at amino acid residue level into several mechanisms employed by nature for the disorder-to-order transition occurring during silk spinning processes, and offers implications in engineering/modulating new biomaterials, via mimicking these natural association pathways.

4.5 Acknowledgement

Authors gratefully acknowledges generous support from NSERC and the University of Ottawa.

CHAPTER 5

Concluding Remarks

The role of protein disorder in the majority of cellular functions is unquestionable. However, there are many challenges to explain the mechanisms behind functional disorder in intrinsically disordered protein, which are distinctly different from the established paradigms for ordered proteins (discussed in Chapter 1). IDP functionality is characterized by highly dynamic conformational fluctuations that exhibit unusual elastic properties and that facilitate intermolecular binding events and triggered structural transitions. These three phenomena, IDP elasticity, disordered-complex formation, and disorder-to-order transitions were studied computationally at atomistic resolution inaccessible to experiments.

In the first study (Chapter 2), structural and mechanical aspects of an elastomeric disordered polypeptide (WGM), inspired by repeat motifs in the domains responsible for the high elasticity of Wheat Gluten protein, were investigated. First, it was revealed that this designed peptide has robust intrinsically disordered characteristics with elastic properties that are qualitatively different from both natively ordered proteins and simple homopolymers of the same size. An attempt was made to use the worm-like chain model to fit the force-extension profile of WGM, but this failed to describe the observed differences in mechanical response occurring across the applied force range. Two main sources of elastic behaviour were identified as dihedral angle deformation (at high extension/force) and hydrophobic residue clustering (at low extension/force). In addition, a structural coil-turn transition occurring locally determined elasticity at intermediate forces/extensions and was influenced by protein sequence.

In the second study (Chapter 3), the variable B-domain in Drp1 protein, a protein responsible for mitochondrial membrane scission, with unknown structure and function was investigated. This domain showed more than 80% irregular structure and did not undergo a significant structural transition even in the presence of a stabilizer osmolyte (TMAO). This contradicts a previously proposed scheme for B domain function, in which it behaves as an auto-inhibitor that is switched off upon making a disorder-to-order transition. Such a mechanism was proposed to facilitate the Drp1 self-oligomerization pathway required for membrane scission. Further analysis indicated that polar TMAO molecules were able to reduce inter-chain electrostatic repulsion, leading to association of B-domain into dimers without folding to an ordered structure. This suggests that modification of the charged residue environments on the B domain surface can induce self-oligomerization, which possibly facilitates Drp1 self-assembly or protein-membrane associations in biologically-relevant conditions. Negatively charged cardiolipin, a lipid that is unique to the mitochondrial membranes that are the tar-

get for DRP1 function may act in way analogous to TMAO molecules in this simulation study, which would explain how Drp1 protein selectively targets the mitochondrial membrane.

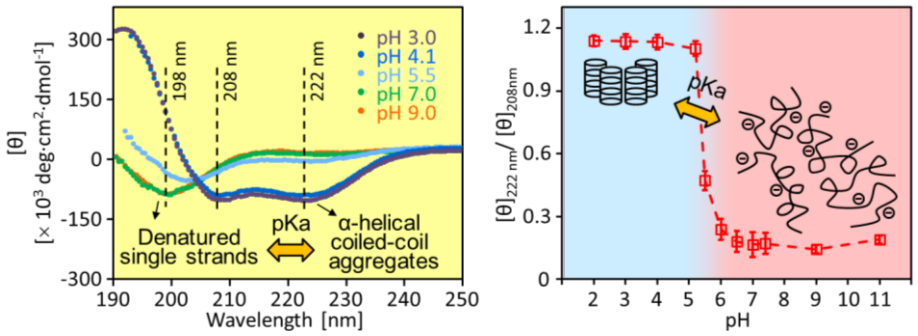
Finally, in the third study (Chapter 4), disorder-to-order transition mechanisms for the assembly of disordered silk-mimetic peptides into beta sheet structures was studied. These silk-mimetic peptides were intended as models for functional domains in full length fibroin silk during the fiber spinning process. Among other physicochemical parameters, such as pH drop and ionic strength, involved in silk spinning, the influence of silk sequence and flow-induced extension/tension on the beta-sheet secondary structure in self-assembly of the silk-like peptides was explored. Tyrosine residues showed high propensity for co-localization via their transient aromatic group interactions, thereby enhancing the inter-chain associations which are a prerequisite for beta-sheet formation. This tyrosine co-localization effect was significantly pronounced for chemically cross-linked tyrosines, a situation occurring in some natural silks. It was also revealed that applied tension stabilizes beta-structure, preferentially for antiparallel beta sheet at higher tensions.

Overall, This study provides substantive insight into various aspects of disordered protein functionality. Based on intuition established in this work, numerous questions are apparent that might motivate future work. One important challenge is to modify existing polymer models to incorporate distinct elastic responses that occur in different force regimes, as well as for heterogeneous sequences with significant proline content. A more specific question would concern how B-domains would be affected by the presence of cardiolipin or GDP in solutions, as is the case in native conditions. A more general question would be about the roles of highly polar solutes on the association behaviour of IDPs with patches of surface charge. It would also be interesting to consider whether residues with bulky phenyl rings play a more general role in templating disorder-order transitions in IDPs, as they seem to in silk based on our results. Further analyses of the kind presented here may shed more light on the role of dynamic conformational disorder in IDPs on their biological functions in cellular processes.

Appendices

APPENDIX A

pH-triggerable conformational switch in alpha-helical coiled-coils



pH-triggerable conformational switch in α-helical coiled-coils

Yinan Lin,[†, #] Mehran Bagheri,[‡, #] Bo An,[†] Qianrui Wang,[†] James L. Harden,[‡,*] and David L. Kaplan[*,†]

[†] Department of Biomedical Engineering, Tufts University, Medford, Massachusetts 02155, United States

[‡] Department of Physics, University of Ottawa, Ontario K1N 6N5, Canada

[#] Contributed equally to this work

[*] Corresponding Authors: david.kaplan@tufts.edu and jharden@uottawa.ca

ABSTRACT

We report the design and characterization of a new α-helical peptide, E4MTA, exhibiting highly-controlled conformational responses to narrow changes in solution pH. This was achieved by introducing along a 4-heptad dimer-forming coiled-coil, E0MTQ, a combination of glutamic acid and alanine substitutions, some of which destabilize the cooperatively-assembled α-helices at neutral and higher pHs, while others promote the α-helix formation at acidic pHs. The resulting trigger-pH-window in E4MTA was narrowed to a range from pH 5.2 to 6.0, as compared to a range from 3.0 to 7.0 in the non-substituted E0MTQ, as revealed by circular dichroism spectroscopy. Moreover, the E4MTA peptide exhibited a propensity to form higher-order coiled-coil aggregates (predominately tetramers), as corroborated by size exclusion chromatography and computer simulation studies. Our findings could expand the dimension of dynamic control over chemically/electrochemically switchable physical crosslinks in peptide-based materials and shed new insight into construction of responsive chimeric proteins.

Desired dynamic responses to various environmental cues are key in the design and development of the next-generation intelligent biomaterials to address the increasingly sophisticated challenges facing the field of biomedical technologies.[1] In recent years, chimeric protein-based hydrogels have emerged as promising materials for tailor-made, multitasking platforms in tissue engineering and regenerative medicine.[2-4] Extensive efforts, therefore, have been devoted to expand the toolkit of stimuli-responsive peptide motifs, ranging from chemically and chemo-enzymatically active sites to supramolecular self-assembled nanostructures.[5]

Among the spectrum of options, α-helical coiled-coils stand out due to their propensity to form super-coiled multi-strands.[6-8] At the amino acid sequence level, classical coiled-coils (C_c) are characteristically comprised of consecutive heptad motifs, denoted as **a-b-c-d-e-f-g**. The amino acid residues at the **a** and **d** positions are predominantly hydrophobic, and pack side-by-side into a hydrophobic core in a "knobs-into-holes" manner.[9-11] The proximal **e** and **g** positions are typically occupied by charged residues, which mediate the intra- and inter-helical electrostatic interactions, and thus modulate the stability of each oligomeric state.[9-11] The residues at the surface exposed **b**, **c**, and **f** positions play a role in the inherent stability of the helical state and also act collectively to organize solvent structure in the vicinity of the coiled-coil assemblies.[11-13] Over the last few decades, synthetic proteins have been successfully engineered with C_c domains to accomplish a variety of functions, including self-assembly of fibrils, hydrogels and biofunctional surfaces, liquid crystallinity, piezoelectricity, modular molecular recognition, and controlled hydrogel erosion.[8,14-23]

Conformational switching behavior between multi-stranded superhelices and single-stranded random coils is an important consideration in the *de novo* design of C_c motifs as pH-reversible triggers. Research efforts have focused on widening the reach of the trigger pHs. Histidine substitutions, for example, were introduced at the conventionally hydrophobic core **d** positions of alternate heptads in a 41-residue coiled-coil peptide, to increase the inter-helical repulsion at the low end of the pH scale and, consequently, triggered formation of α-helical fibers at pH ~ 6.0, near the pK_a of the free imidazole side chain of histidine.[23,24] Moreover, the pK_a values of the amino acid side chain groups have been further modulated by selected conformational attributes, such as peptide chain folding, inter-chain association and the establishment of a local hydrophobic pocket. For instance, when glutamic acid residues (with a free side chain carboxyl group pK_a of 4.3), were placed in the peripheral **e** and **g** positions, the conformational switching in the reported C_c peptides shifted to a pH between 5.5 and 6.0.[25] Even at the surface **b** and **c** positions, charged residues, e.g., lysine and glutamic acid, were found to affect the pH-dependent stability of C_c motifs.[26]

The pH-induced conformational responses in modular C_c domains have been shown to be rapid and reversible, and hence suitable for dynamic sol-gel switching.[27-31] Emerging applications, such as electrochemical patterning of hydrogel films controlled by local pH gradients,[32-34] require tight spatial control of molecular properties. For electrochemical control of conformational switching, it is highly desirable to minimize the pH span of the transition, an aspect that has not been well studied. In this Communication, we report our recent progress in fine-tuning a narrow trigger pH window in a modular C_c polypeptide by using a novel strategy combining amino acid substitutions that stabilize the coiled-coil structures at low pHs and destabilize them at higher pHs. We present the detailed design, modeling and characterization of a peptide (denoted E4MTA) that is capable of completing the coil-helix switching within a narrow pH range between 5.2 and 6.0. This is in contrast to other reported pH switchable systems, for which the transition ranges generally span 2 to 3 pH units.

The sequence design of E4MTA is comprised of modified sequences of previously reported dimerization motifs utilized in de novo fibers and hydrogels[11,14]. Our starting point was a four-heptad helical dimerization motif, E0MTQ (Table 1), inspired by these dimeric coiled-coil systems.[14] A set of four glutamic acid substitutions in E0MTQ were implemented in E4MTA: one at the peripheral **e** position in the first heptad repeat (K5E), two at the surface-exposed **b** and **c** positions in the two middle repeats (Q9E and Q17E), and one at the core **d** position in the fourth heptad repeat (L25E). Notably, alanine residues were also introduced to replace all of the glutamine residues at the surface-exposed **f** positions, in

order to enhance coiled-coil stability at low pH[12,13]. In contrast, the glutamic acid substitution in the hydrophobic core **d** position was chosen to destabilize coiled-coil assemblies at elevated pH.

Circular dichroism (CD) spectra were collected from 190 to 250 nm for the two polypeptides (both 30 μM in aqueous buffers, described in the Supplemental Materials section, at 20°C) (Figure 1a-1b) over a pH range from 2.0 to 11.0. The α-helical structures are characterized by two negative dichroic minima at 208 and 222 nm, and higher-order coiled-coil oligomerization is reflected by a minimum at 222 nm that is more negative than the one at 208 nm.[35] The pH-dependence of the ratio of ellipticities at 222 and 208 nm, $\rho=[\theta]_{222nm}/[\theta]_{208nm}$, are summarized in Figure 1c for comparison. The original E0MTQ peptide displayed a broad transition zone between α-helices at pH 3.0 and denatured single strands at pH 7.0, the later indicated by an intense negative band at ~ 200 nm (Figure 1a). However, CD wavelength scans of the modified E4MTA peptide (Figure 1b) at various pHs revealed a contrastingly sharp helix-coil transition occurring in the pH range between pH 5.2 and 6.0, demonstrating higher pH sensitivity and selectivity. When the polypeptides were switched into their denatured random coil conformations above the trigger points, $pH > pH_{Trigger}$, the ratio ρ dropped significantly, suggesting a loss of oligomerization. For the case of E4MTA, this drop in ρ (from $\rho\approx1.1$ to 0.24) across the trigger point ($pH_{Trigger} \approx 5.3$) was sharper and more precipitous than for the control peptide E0MTQ, which exhibited a smooth decrease from $\rho\approx0.9$ to 0.25 across a relatively broad pH range from pH 3 to pH 7. Moreover, the values of ρ at low pH for the two peptides suggest an increase in the coiled-coil stability in E4MTA ($\rho\approx1.1$) as compared with E0MTQ ($\rho\approx0.9$) (Figure 1c).

The enhanced coiled-coil stability in E4MTA at low pH was further verified by differential scanning calorimetry (DSC) for 5 mM peptide solutions in 10 mM citric buffer at pH 3.5 (Figure 1d). The original E0MTQ peptide exhibited a broad denaturation transition between between 20 and 75°C, as indicated by a symmetric melting peak centered around 49°C. In contrast, the E4MTA peptide had a higher temperature denaturation transition occurring over a relatively narrower temperature range between 45 and 80°C, as indicated by an asymmetric melting peak centered around 74°C. The asymmetric shape of the E4MTA peptide thermal transition suggests the existence of several oligomeric states of this peptide, whereas the original E0MTQ peptide is a putative dimer former.

The existence of higher order oligomers in E4MTA was verified in analytical size exclusion chromatography (SEC) studies conducted at pH 4 on 150 μM peptide solutions in 10 mM acetate buffer (Figure 2). The SEC data shows four clear oligomeric states. Specifically, deconvolution of the SEC data indicated the existence of a dominant tetramer (39%) in coexistence with monomeric E4MTA (33%) and dimeric (14%) and trimeric aggregates (15%) of E4MTA in acidic pH conditions.

To further understand the coiled-coil oligomerization at the molecular level, all atom molecular dynamics (MD) simulations, conducted in 10 mM NaCl solutions under NPT conditions at 300 K and 1 atm., were performed using the NAMD simulation package[36]. The CHARMM22 force field [37] and the TIP3P model [38] were used for the peptide and water molecules, respectively. Dimeric, trimeric, and tetrameric configurations of E4MTA were constructed from its helical monomer conformation. In particular, aggregates with hydrophobic cores of the classical C_c type comprised of the **a** and **d** hydrophobic residue faces (denoted HP), hydrophobic cores comprised of alanine residues from the **f** positions (denoted ALA), and mixed cores comprised of **a**-**d** hydrophobic residue faces in contact with alanines from **f** positions (denoted ALA-HP) were initially investigated.

Each oligomeric state was equilibrated for at least 100 ns. Dimeric, trimeric, and tetrameric oligomers of E4MTA with the classical HP-cores were found to be stable in low pH conditions (as were, to a lesser degree, dimers with mixed ALA-HP cores). The changes in free energies associated with the formation of oligomers with HP-cores were then obtained from multiple trials using the free energy perturbation (FEP) method [39,40]. The calculated free energy of association on a per peptide basis was lowest for HP-based tetramers (-13.2 ± 2.2 kCal/mol), followed by trimers (-10.3 ± 2.3 kCal/mol), and dimers (-4.6 ± 2.2 kCal/mol), in qualitative agreement with the SEC data.

Several competing physical effects appear to play roles in the preference for tetramers of E4MTA at low pH (Figure 3). The Glu in the core **d** position at residue 25, introduced to suppress oligomerization above the pH trigger point, also plays a destabilizing role at low pH in dimeric and trimeric oligomers,

due to steric hindrance of the Glu side chain in the hydrophobic pocket; this Glu side chain appears to be well accommodated in the tetrameric state of association (Figure 3c). On the other hand, the Glu residues at the **b** and **c** positions interact with Glu and Lys residues at **e** and **g** positions to stabilize the tetrameric state at low pH (Figure 3b). However, above the trigger pH, the negatively-charged Glu at residue 25 can no longer be accommodated in the pocket, leading to loss of oligomer stability and the concomitant suppression of the alpha-helical conformation of individual peptides by intra-molecular electrostatic repulsion between the many Glu residues in the solvent exposed faces.

In conclusion, a new polypeptide module E4MTA has been reported that enables unprecedented control of the pH switching behavior between the tetrameric coiled-coil and disordered states. The acid-triggered, reversible coil-helix transition was fine-tuned to a sharply defined pH range between 5.2 and 6.0 through judiciously chosen Glu and Ala residues that stabilze the coiled-coil at low pH while destabilizing it at high pH. The formation of higher order oligomers (primarily tetramers) below the pH trigger point is a feature that enhances its utility for use as a triggering moiety for biomaterials applications (e.g. as a pH-sensitive cross-linking domain in hydrogel forming peptides). More broadly speaking, such tight pH control over secondary and quaternary structure will facilitate development of spatially and temporally patterned biomaterials that may be externally-controlled electrochemically, for instance by embedded patterned electrodes.

ASSOCIATED CONTENT

Supporting Information
Materials and methods, MALDI-TOF MS, variable-temperature CD spectra of the *de novo* designed C_c candidate peptides. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Authors:
david.kaplan@tufts.edu and jharden@uottawa.ca

Author Contributions:
Yinan Lin and Mehran Bagheri contributed equally to this work.

Notes
The authors declare no competing financial interests.

ACKNOWLEDGEMENT

The authors gratefully acknowledge funding from the NIH P41 EB002520 and U01 EB014976 and from the NSERC CREATE and Discovery grant programs.

REFERENCES

(1) Wojtecki, R. J.; Meador, M. A.; Rowan, S. J. *Nat. Mater.* **2011**, *10*, 14-27.
(2) Scholler, J.; Brady, T. L.; Binder-Scholl, G.; Hwang, W. T.; Plesa, G.; Hege, K. M.; Vogel, A. N.; Kalos, M.; Riley, J. L.; Deeks, S. G.; Mitsuyasu, R. T.; Bernstein, W. B.; Aronson, N. E.; Levine, B. L.; Bushman, F. D.; June, C. H. *Sci. Transl. Med.* **2012**, *4*, 132ra53.
(3) Nuccitelli, A.; Cozzi, R.; Gourlay, L. J.; Donnarumma, D.; Necchi, F.; Norais, N.; Telford, J. L.; Rappuoli, R.; Bolognesi, M.; Maione, D.; Grandi, G.; Rinaudo, C. D. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 10278-10283.
(4) Ansari, S.; Moshaverinia, A.; Pi, S. H.; Han, A.; Abdelhamid, A. I.; Zadeh, H. H. *Biomaterials* **2013**, *34*, 10191-10198.
(5) Kushner, A. M.; Guan, Z. B. *Angew. Chem., Int. Ed.* **2011**, *50*, 9026-9057.
(6) Fletcher, J. M.; Harniman, R. L.; Barnes, F. R. H.; Aimee L. Boyle1; Collins, A.; Mantell, J.; Sharp, T. H.; Antognozzi, M.; Booth, P. J.; Linden, N.; Miles, M. J.; Sessions, R. B.; Verkade, P.; Woolfson, D. N. *Science* **2013**, *340*, 595-599.
(7) Landschulz, W. H.; Johnson, P. F.; Mcknight, S. L. *Science* **1988**, *240*, 1759-1764.
(8) Petka, W. A.; Harden, J. L.; McGrath, K. P.; Wirtz, D.; Tirrell, D. A. *Science* **1998**, *281*, 389-392.
(9) Crick, F. H. C. *Acta Crystallogr.* **1953**, *6*, 689-697.
(10) Lupas, A. *Trends Biochem. Sci.* **1996**, *21*, 375-382.
(11) Walshaw, J.; Woolfson, D. N. *J. Struct. Biol.* **2003**, *144*, 349-361.
(12) O'Neil, K.T.; Degrado, W.F. *Science* **1990**, *250*, 646-651.
(13) Dahiyat, B. I.; Gordon, D. B.; Mayo, S.L. *Protein Sci.* **1997**, *6*, 1333-1337.
(14) Banwell, E. F.; Abelardo, E. S.; Adams, D. J.; Birchall, M. A.; Corrigan, A.; Donald, A. M.; Kirkland, M.; Serpell, L. C.; Butler, M. F.; Woolfson, D. N. *Nat. Mater.* **2009**, *8*, 596-600.
(15) Mi, L.; Fischer, S.E.; Chung, B.; Sundelacruz, S.; Harden, J. L. *Biomacromolecules* **2006**, *7*, 38-47.
(16) Fischer, S. E.; Liu, X.; Mao, H.-Q.; Harden, J. L. *Biomaterials* **2007**, *28*, 3325--3337.
(17) Yu, S. J. M.; Conticello, V. P.; Zhang, G. H.; Kayser, C.; Fournier, M. J.; Mason, T. L.; Tirrell, D. A. *Nature* **1997**, *389*, 167-170.
(18) Zhang, G. H.; Fournier, M. J.; Mason, T. L.; Tirrell, D. A. *Macromolecules* **1992**, *25*, 3601-3603.
(19) Ghosh, I.; Hamilton, A. D.; Regan, L. *J. Am. Chem. Soc.* **2000**, *122*, 5658-5659.
(20) Magliery, T. J.; Wilson, C. G. M.; Pan, W. L.; Mishler, D.; Ghosh, I.; Hamilton, A. D.; Regan, L. *J. Am. Chem. Soc.* **2005**, *127*, 146-157.
(21) Shen, W.; Zhang, K. C.; Kornfield, J. A.; Tirrell, D. A. *Nat. Mater.* **2006**, *5*, 153-158.
(22) Farrar, D.; Ren, K. L.; Cheng, D.; Kim, S.; Moon, W.; Wilson, W. L.; West, J. E.; Yu, S. M. *Adv. Mater.* **2011**, *23*, 3954-3958.
(23) Zimenkov, Y.; Dublin, S. N.; Ni, R.; Tu, R. S.; Breedveld, V.; Apkarian, R. P.; Conticello, V. P. *J. Am. Chem. Soc.* **2006**, *128*, 6770-6771.
(24) Wada, K.; Mizuno, T.; Oku, J.; Tanaka, T. *Protein Pept. Lett.* **2003**, *10*, 27-33.
(25) Kohn, W. D.; Monera, O. D.; Kay, C. M.; Hodges, R. S. *J. Biol. Chem.* **1995**, *270*, 25495-25506.
(26) Vu, C.; Robblee, J.; Werner, K. M.; Fairman, R. *Protein Sci.* **2001**, *10*, 631-637.
(27) Hammes, G. G.; Roberts, P. B. *J. Am. Chem. Soc.* **1969**, *91*, 1812-1816.
(28) De Sancho, D.; Best, R. B. *J. Am. Chem. Soc.* **2011**, *133*, 6809-6816.
(29) Huang, C. Y.; Getahun, Z.; Wang, T.; DeGrado, W. F.; Gai, F. *J. Am. Chem. Soc.* **2001**, *123*, 12111-12112.
(30) Bredenbeck, J.; Helbing, J.; Kumita, J. R.; Woolley, G. A.; Hamm, P. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 2379-2384.
(31) Tucker, M. J.; Abdo, M.; Courter, J. R.; Chen, J. X.; Brown, S. P.; Smith, A. B.; Hochstrasser, R. M. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 17314-17319.
(32) Leisk, G. G.; Lo, T. J.; Yucel, T.; Lu, Q.; Kaplan, D. L. *Adv. Mater.* **2010**, *22*, 711-715.
(33) Kojic, N.; Panzer, M. J.; Leisk, G. G.; Raja, W. K.; Kojic, M.; Kaplan, D. L. *Soft Matter* **2012**, *8*, 6897-6905.
(34) Lin, Y. N.; Xia, X. X.; Shang, K.; Elia, R.; Huang, W. E.; Cebe, P.; Leisk, G.; Omenetto, F.; Kaplan, D. L. *Biomacromolecules* **2013**, *14*, 2629-2635.
(35) Marsden, H. R.; Korobko, A. V.; van Leeuwen, E. N. M.; Pouget, E. M.; Veen, S. J.; Sommerdijk, N. A. J. M.; Kros, A. *J. Am. Chem. Soc.* **2008**, *130*, 9386-9393.

(36) Phillips, J.C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.D.; Kal, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
(37) 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.; Wirkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
(38) Jorgensen, W.L.; Chandrasekhar, J.; Madura, J. D.; Impey, R.W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
(39) Kollman, P.; *Chem. Rev.* **1993**, *93*, 2395–2417.
(40) Pohorille, A.; Jarzynski, C.; Chipot, C. *J. Phys. Chem. B* **2010**, *114*, 10235–10253.
(41) Humphrey, W.; Dalke, A.; Schulten, K. *J. Molec. Graphics*, **1996**, *14*, 33-38.

Table 1. The amino acid sequences of two peptides: E0MTQ and E4MTA.

C_c peptides	Peptide sequences			
Heptad repeat	***a b c d e f g***	***a b c d e f g***	***a b c d e f g***	***a b c d e f g***
E0MTQ	IQQLKQE	NQQLKQE	IQQLKQE	IQQLKQE
E4MTA	IAALEAE	NEALKAE	IAELKAE	IAAEKAE

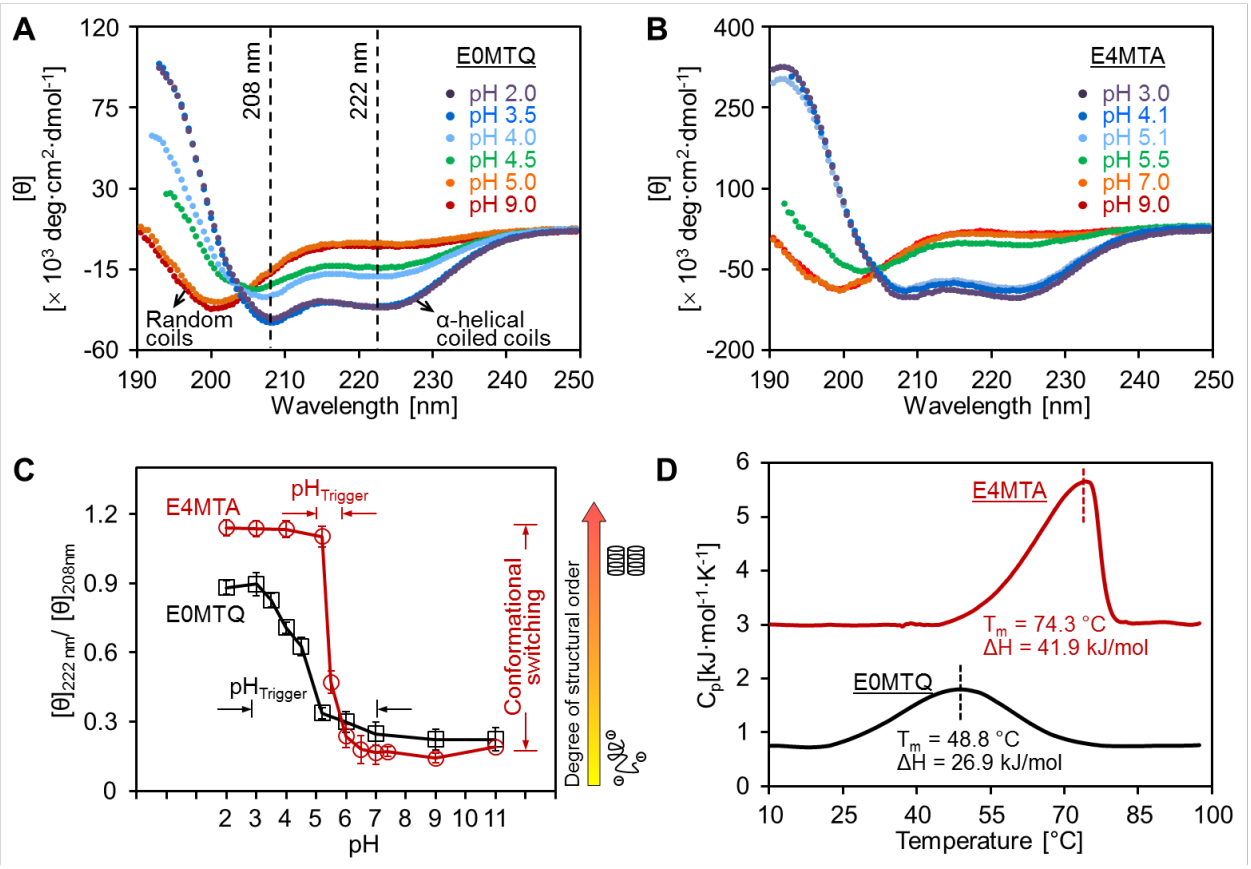


Figure 1. Variable-pH CD spectroscopy of (A) E0MTQ, and (B) E4MTQ (both 30 µM in aqueous buffers). (C) Ratios of mean residue ellipticities at 222 and 208 nm of the two polypeptides, E0MTQ (black squares) and E4MTA (red circles), as a function of pH. The plot reveals a sharp conformational transition between α-helices and denatured conformations in E4MTA around pH 5.3, denoted as $pH_{Trigger}$, as opposed to a gradual switch shown in E0MTQ with a midpoint around pH 4.2. (D) DSC thermograms of the two polypeptides (bottom to top): E0MTQ (black) and E4MTA (red) (both 5 mM in a 10 mM citric acid/sodium citrate buffer at pH 3.5)

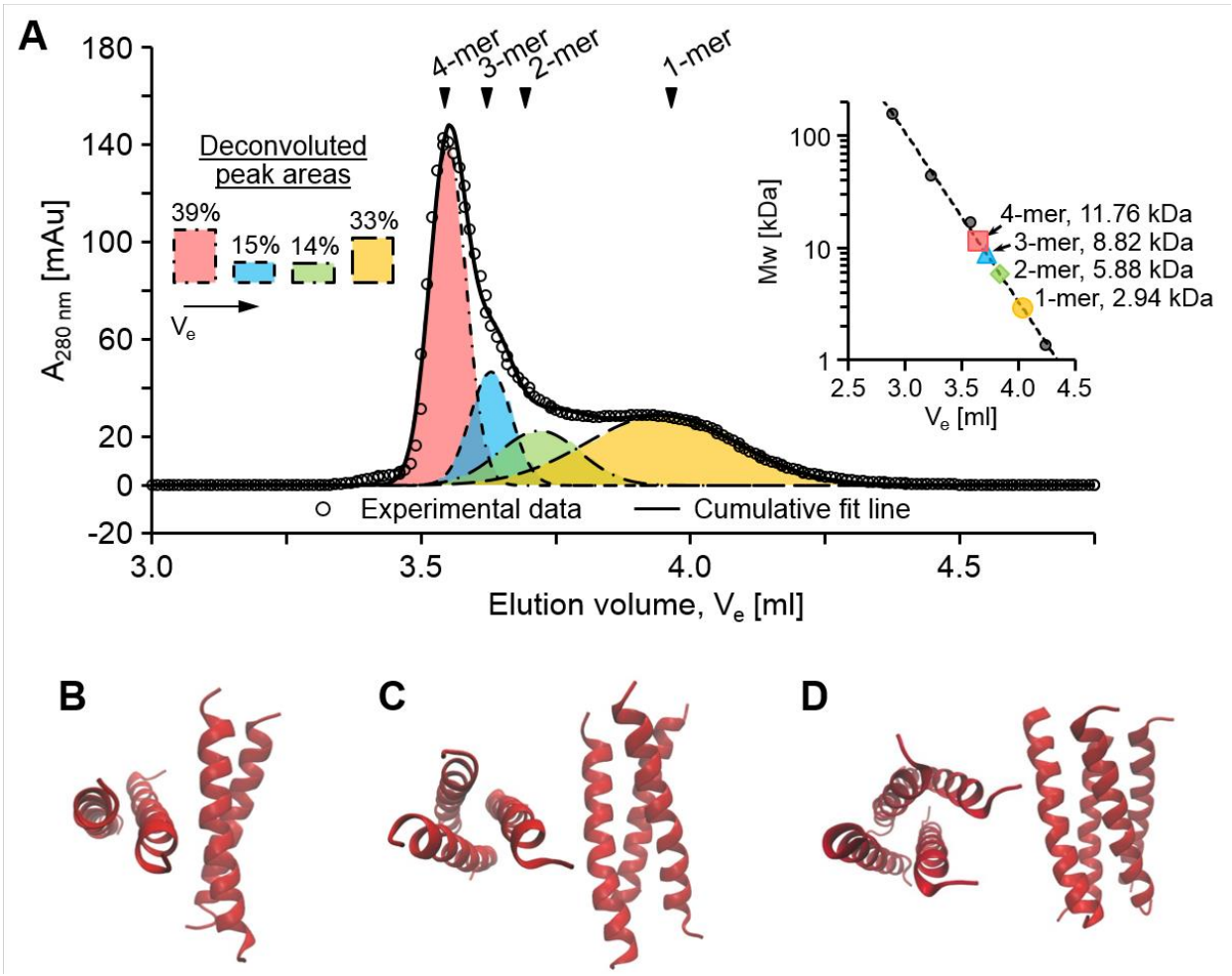


Figure 2. (A) Size exclusion chromatography elution profile of E4MTA and Gaussian fitting peaks corresponding to unimers (yellow), dimers (green), timers (blue), and tetramers (pink). The polypeptide was prepared in an acetate stock buffer (10 mM, pH 7.4) at a concentration of 150 μM, and then injected into the column buffered at pH 4.0 with 10 mM acetate buffer at 4 °C. Inset shows the elution data plotted semi-logarithmically along with the calibration standards. (B-D) Top and side views of schematic dimeric, trimeric, and tetrameric coiled coil E4MTA peptide assemblies.

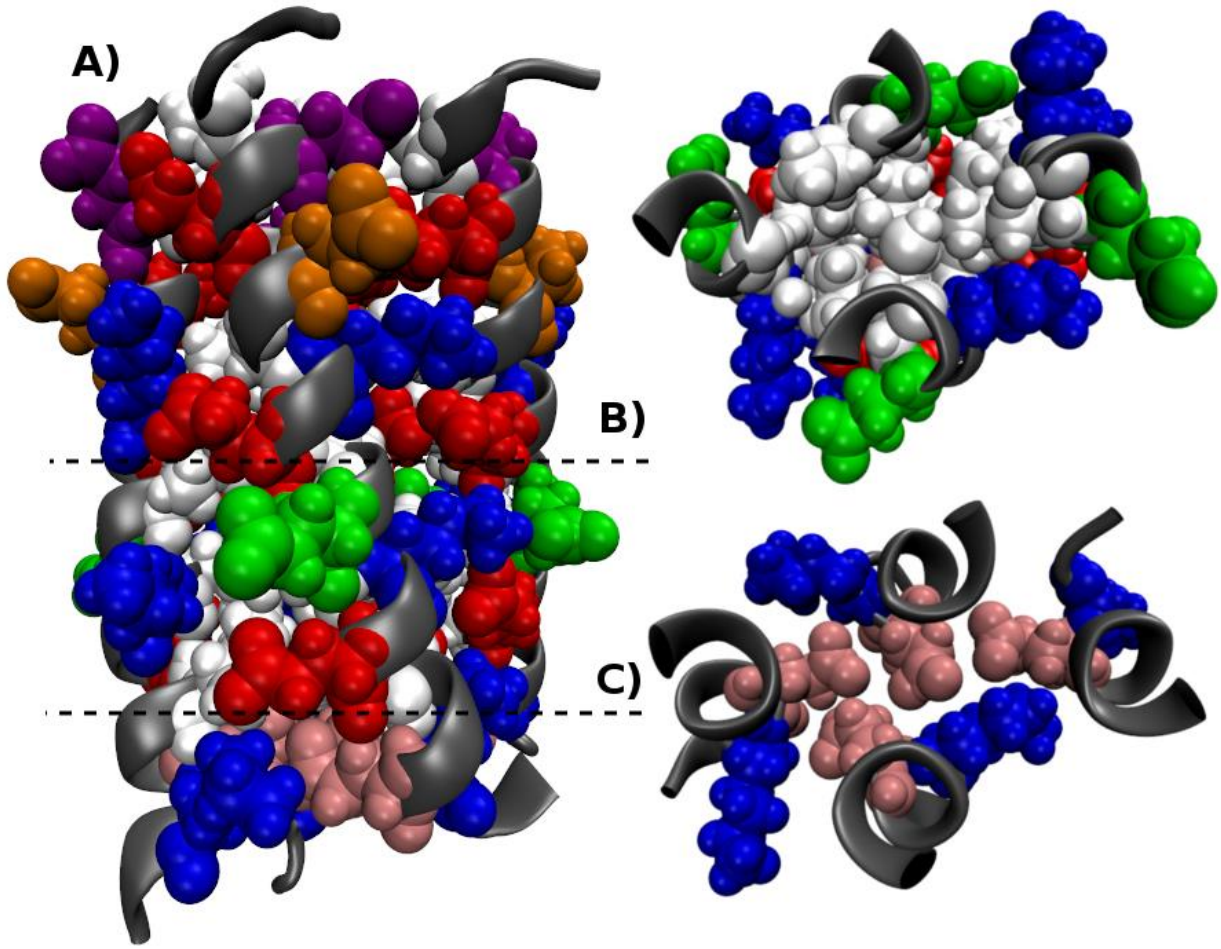


Figure 3. (A) Snapshot of a side-view of an equilibrated E4MTA tetramer, showing Glu residues in the **b** (orange), **c** (green), and **d** (pink) positions, along with Glu (red) and Lys (blue) residues in the classical **e** and **g** positions; (B) cross sectional view of a cut through the E4MTA tetramer, showing the arrangement of the Glu substitutions in the **c** (green) positions; (C) cross sectional view of a cut through the E4MTA tetramer, showing the arrangement of the Glu substitutions in the **d** (pink) hydrophobic core positions at residue 5. The VMD package was used to produced these images[41]

SUPPORTING INFORMATION

pH-triggerable conformational switch in α-helical coiled-coils

Yinan Lin,[†, #] Mehran Bagheri,[‡, #] Bo An,[†] Qianrui Wang,[†] James L. Harden,[‡,*] and David L. Kaplan[*,†]

[†] Department of Biomedical Engineering, Tufts University, Medford, Massachusetts 02155, United States

[‡] Department of Physics, University of Ottawa, Ontario K1N 6N5, Canada

[#] Contributed equally to this work

[*] Corresponding Authors: david.kaplan@tufts.edu and jharden@uottawa.ca

TABLE OF CONTENTS

Materials

All peptides (Table 1 & Table S1) (all > 98 %, HPLC) were purchased from GL Biochem (Shanghai, China). Citric acid (≥ 99.5%), sodium citrate dihydrate (≥ 99%), sodium acetate (CH_3COONa, ≥ 99%), acetic acid (CH_3COOH, ≥ 99%), 4-(2-hydroxyethyl)piperazine-1-ethanesulfonic acid (HEPES, ≥ 99.5%), hydrochloric acid (HCl, 37%), tris(hydroxymethyl)aminomethane (Tris base, ≥ 99.8%), sodium phosphate dibasic (Na_2HPO_4, ≥ 99%), sodium bicarbonate ($NaHCO_3$, ≥ 99.5%), and sodium hydroxide (NaOH, ≥ 98%) were purchased from Sigma-Aldrich (St. Louis, MO). Phosphate buffered saline (PBS, 10X, pH 7.4) and Ultrapure™ water were obtained from Life Technologies (Carlsbad, CA). All chemicals were used without further purification.

Methods

MALDI-TOF mass spectrometry

MALDI matrix was prepared by saturating sinapinic acid (Sigma) in solution contains 50% (v/v) acetonitrile and 0.3% (v/v) trifluoroacetic acid. A 6 µL aliquot of 1 mg/mL sample was mixed with 24 µL of matrix, 1 µL solution was plated onto a 96 spot target plate and allowed to dry. MALDI-TOF mass spectra (Figure S1) were acquired on a Microflex LT system (Bruker Corporation, Billerica, MA) with 50% laser intensity using standard RN (reflectron negative) pepMix method provided by the software.

Preparation of buffer stocks and peptide solutions

Different buffers (Table S2) were used to study how the pH affects the conformation of the peptides: six 1M citric acid/sodium citrate buffers of pH 3.0, 4.0, 4.5, 5.5, 6.0 and 6.5, one 1M sodium acetate/acetic acid buffer of pH 5.1, one 1M HEPES buffer of pH 7.0, one 0.1M PBS buffer of pH 7.4, three 1M Tris/HCl buffers of pH 8.0, 8.5 and 9.0, three 0.5M phosphate buffers of pH 9.5, 10 and 10.5, and one 1M sodium bicarbonate/sodium hydroxide buffer of pH 11.0. All peptide solutions were prepared by dilution of one of the concentrated buffer stocks to an ionic strength of 10 mM using Ultrapure™ water. The peptide concentration was determined by the absorbance measured at 214 nm and calculated using the molar extinction coefficient: ε_{214} = 2,200 $cm^{-1}M^{-1}$ per peptide bond as reported by Persikov *et al*.[S1]

Variable-pH Circular Dichroism (CD) spectroscopy

The CD measurements were carried out in 1 mm path length quartz cuvettes using an Aviv model 62DS spectrophotometer equipped with a Peltier temperature controller (Aviv Biomedical, Lakewood, NJ). The peptides were dissolved in each of the sixteen buffers (10 mM) at a concentration of 30 µM, and allowed to equilibrate at 20°C for 30 min prior to CD experiments. Each spectrum (Figure 1a and b) was then obtained at 20°C in the 190 ~ 260 nm range (but only data from 190 ~ 250 nm is shown) at a resolution of 0.5 nm and at a scanning speed of 50 nm/min. Baselines recorded using the same buffer and cuvette were subtracted from the data. The trigger or switch pH, denoted as $pH_{Trigger}$, where a helix-to-random conformational transition occurred, was determined for each peptide from the pH dependence of the mean residue ellipticity ratios at 208 nm and at 222 nm (Figure 1c).

Variable-temperature CD spectroscopy

Variable-temperature CD experiments were performed on E4MTA at pH = 3.0, 4.0 and 5.1 (all lower than the $pH_{Trigger}$ = 5.3), to check the thermal stability of the coiled coil structures. Spectra (Figure S2) were collected from 190 to 260 nm (but only data from 200 ~ 245 nm was shown) every 10°C between 20 and 90°C with an equilibration time of 20 min at each desired temperature. All other settings were the same as in the previous section. The melting temperatures, T_m, were defined as the temperature at which half of the coiled-coils dissociated as determined from the thermal dependence of the mean residue ellipticity at 222 nm (Figure S3).

Differential Scanning Calorimetry (DSC)

DSC experiments (Figure 1d) were performed on a NANO DSC II Model 6100 (Calorimetry Sciences Corp, Lindon, UT). Peptide samples were dissolved in citrate buffer pH 3.5 to a concentration of 5 mM. The solutions were dialyzed against the same buffer before measurement to collect the dialyzed buffer as reference in the experiment. Sample solutions were loaded at 0°C into the cell and heated at a rate of 0.1°C/min till 100°C. Data points were plotted in MicroCal Origin 6.0. T_m and enthalpy were calculated based on peak of curve and area under the curve respectively.

Size exclusion chromatography

The pH-triggerable aggregation behaviors of the E4MTA peptide was examined by size exclusion chromatography (Figure 2) using an ÄKTAmicro chromatography system (GE Healthcare, Uppsala, Sweden) equipped with a Yarra SEC-3000 column (Phenomenex, Torrance, CA) in PBS (pH 7.4) and sodium acetate buffer (pH 4.0), respectively, at 4°C. Standard calibration of molecular weights was done with a Gel Filtration Calibration Standard sample (Bio-Rad, Hercules, CA).

Molecular dynamics simulations

Initial helical configurations of the E4MTA sequence and of putative coiled-coil oligomers of helical E4MTA monomers were generated using a python-package developed in our group. Coiled-coils with hydrophobic cores of three types were investigated: the classical C_c type comprised of the **a** and **d** hydrophobic residue faces (denoted HP), hydrophobic cores comprised of alanine residues from the **f** positions (denoted ALA), and mixed cores comprised of **a-d** hydrophobic residue faces in contact with alanines from **f** positions (denoted ALA-HP). In all cases, the relative orientations of E4MTA monomers in these coiled-coil assemblies were chosen to favor intermolecular electrostatic interactions between charged residues in the **e** and **g** heptad positions. These assemblies were solvated in TIP3P water[S3] with 10 mM NaCl in a box with periodic boundary conditions and boundaries that extend 15 angstroms from the polypeptide assembly surface in each direction, to avoid first neighbor simulation box interactions. All simulations were carried out with a time-step of 2 fs using NAMD 2.9[S4] with the CHARMM22 force field[S5] for the polypeptides. Protonated glutamic acid residues emulated low pH solution conditions. A Langevin thermostat and barostat were employed under NPT conditions at 300 K and 1 atm. The cut-off for van der Waals interactions was 12 angstroms and electrostatic interactions were computed using the particle-mesh Ewald algorithm. At least 100 ns equilibration was performed for all configurations studied, without any restraints. The most stable conformations in the last 20 ns were adopted for Free energy calculations, as described below

Free energy calculations

Free energy difference calculations were carried out using the Alchemical Free Energy Perturbation[S6] method, as implemented in the NAMD software package. In Free Energy Perturbation (FEP) methods, free energy difference between two states, A and B, is calculated from

$$\Delta G_{A\to B} = -k_B T \sum_{i=1}^{N} \ln \left\langle \exp\left([H(\lambda_i) - H(\lambda_{i+1})]/k_B T \right) \right\rangle_{\lambda_i} \qquad (1)$$

where $H(\lambda) = (1 - \lambda)H_A + \lambda H_B$, with H_A and H_B representing the Hamiltonian of the initial ($\lambda = 0$) and final ($\lambda = 1$) states. Using thermodynamic integration[S7], the free energy difference between two states, A and B, is obtained by integrating $H(\lambda)$ with respect to λ, from $\lambda = 0$ to $\lambda = 1$. We used this method to estimate the free energy of association for dimeric, trimeric, and tetrameric coiled-coil aggregates by comparing the associated state to fully solvated, dissociated states of the helical peptide. To do so, stable conformations of the helical peptide and its assemblies were chosen from configurations in the last 20 ns of the equilibration simulations described above, and multiple FEP simulations with restrained peptide backbones were conducted under the same conditions as in the equilibration simulations. These were used, along with the computed free energy cost of the backbone restraints, to compare the free energies of association on a per chain basis between associated states with different pockets (e.g. dimers with HP vs ALA pockets) and different states of association (dimers, trimers, and tetramers with HP pockets).

Table S1. Detailed characteristics of C_c peptide candidates. The unmutated sequence E0MTQ was designed based on a family of hydrogelating self-assembling fibers (hSAFs) peptides as previously reported by Banwell *et al*.[S2]

	Cc candidates	Amino acid sequence	# of residues	Theoretical Mw [kDa]	Theoretical pI
1	E0MTQ	IQQLKQENQQLKQEIQQLKQEIQQLKQE	28	3.49	7.17
2	E4MTA	IAALEAENEALKAEIAELKAEIAAEKAE	28	2.94	4.03

Table S2. List of buffer stocks. All peptide solutions were prepared by dilution of the concentrated stock solutions to an ionic strength of 10 mM.

#	pH	Buffering system	Ionic strength
1	3.0	Citric acid/sodium citrate	1 M
2	4.0	Sodium acetate/acetic acid	1 M
3	4.5	Citric acid/sodium citrate	1 M
4	5.1	Sodium acetate/acetic acid	1 M
5	5.5	Citric acid/sodium citrate	1 M
6	6.0	Citric acid/sodium citrate	1 M
7	6.5	Citric acid/sodium citrate	1 M
8	7.0	HEPES	1 M
9	7.4	PBS	0.1 M
10	8.0	Tris/HCl	1 M
11	8.5	Tris/HCl	1 M
12	9.0	Tris/HCl	1 M
13	9.5	Disodium hydrogen phosphate/sodium hydroxide	0.5 M
14	10.0	Disodium hydrogen phosphate/sodium hydroxide	0.5 M
15	10.5	Disodium hydrogen phosphate/sodium hydroxide	0.5 M
16	11.0	Sodium bicarbonate/sodium hydroxide	1 M

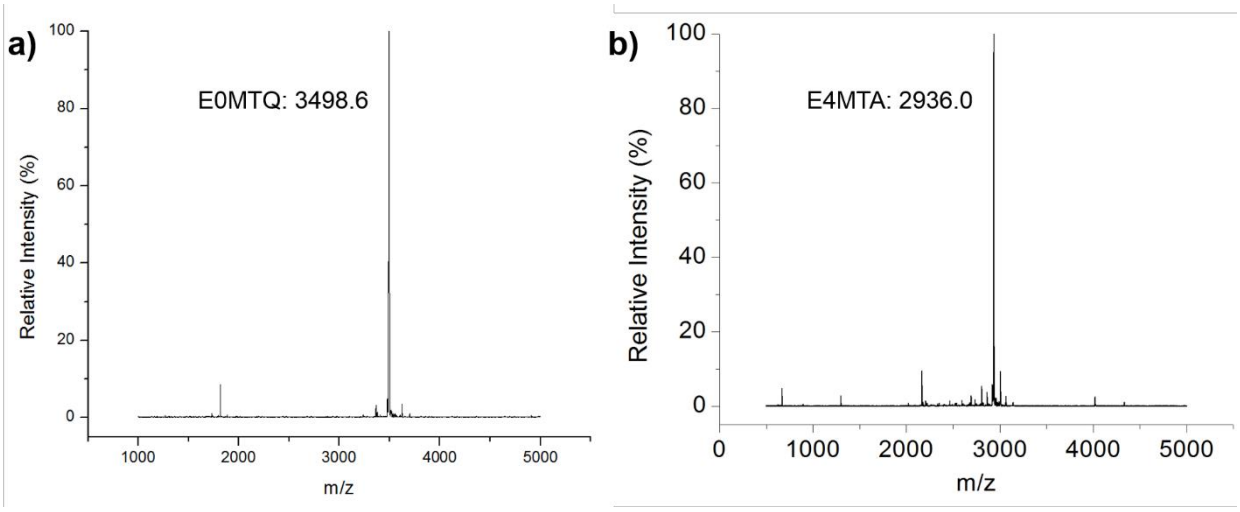


Figure S1. MALDI-TOF MS analysis of the (a) E0MTQ, and (b) E4MTA polypeptides.

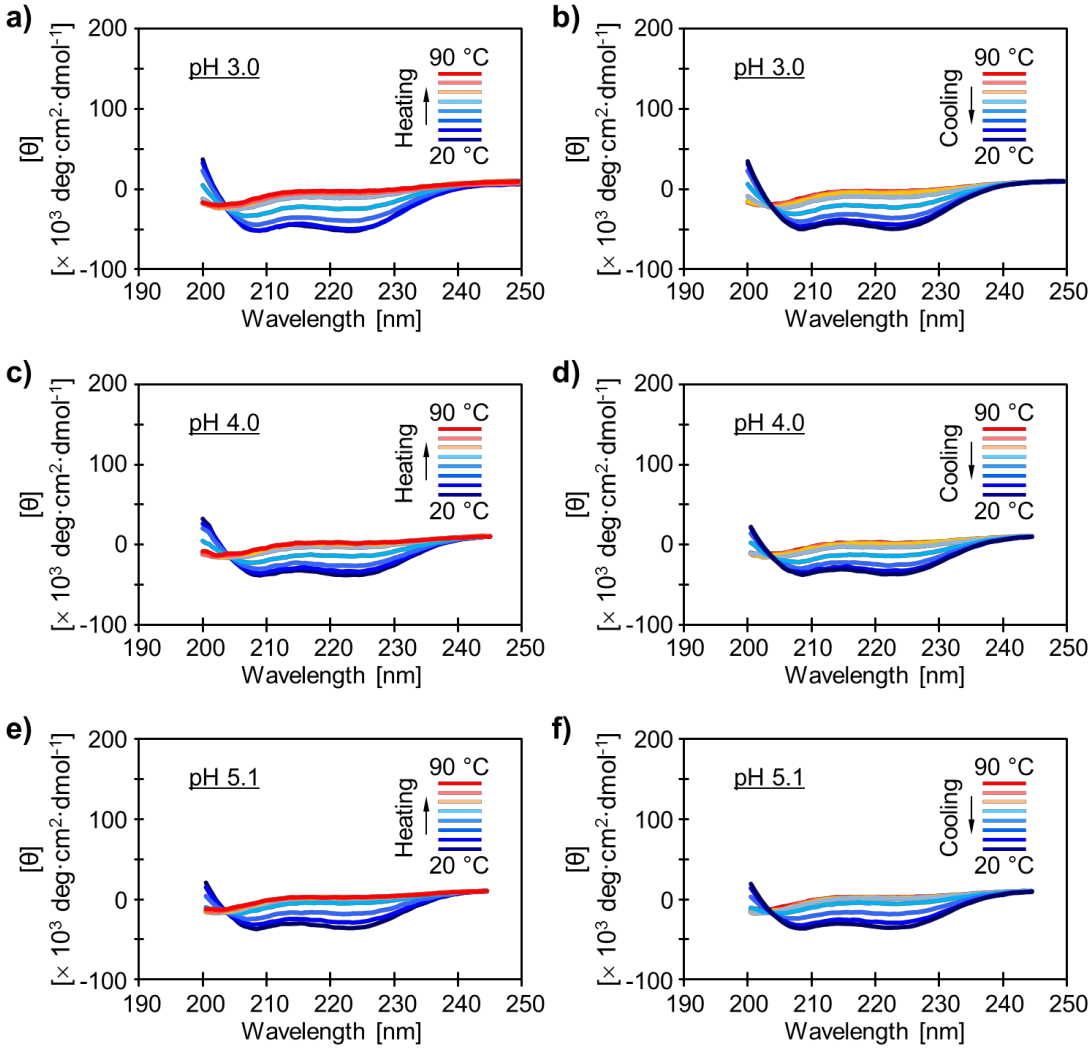


Figure S2. Thermal cycle CD spectra of E4MTA (30 μM) between 20 and 90°C at pH = 3.0 (a-b), 4.0 (c-d) and 5.1 (e-f), respectively.

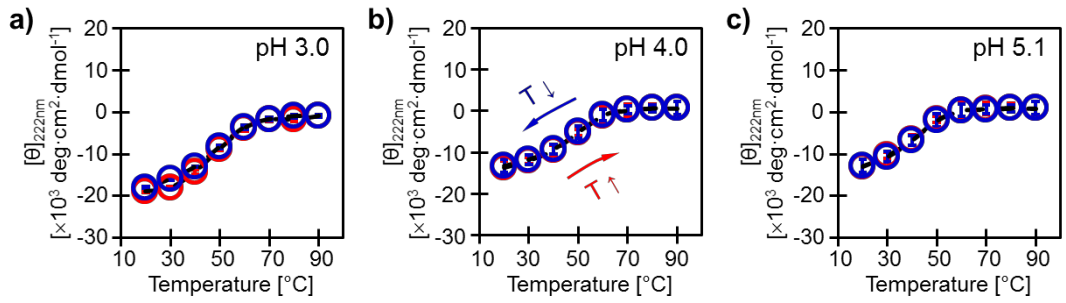


Figure S3. Thermal reversibility of the conformational switch in E4MTA as shown from the temperature-dependent mean residue ellipticity at 222 nm.

References

[S1] Persikov, A. V.; Xu, Y.; Brodsky, B. *Protein Sci.* **2004**, *13*, 893.

[S2] Banwell, E.F.; Abelardo, E.S.; Adams, D.J.; Birchall, M.A.; Corrigan, A.; Donald, A.M.; Kirkland, M.; Serpell, L. C.; Butler, M. F.; Woolfson, D. N. *Nat. Mater.* **2009**, *8*, 596.

[S3] Jorgensen, W.L.; Chandrasekhar, J.; Madura, J. D.; Impey, R.W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

[S4] Phillips, J.C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.D.; Kal, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

[S5] 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.; Wirkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

[S6] Pohorille, A.; Jarzynski, C.; Chipot, C. *J. Phys. Chem. B* **2010**, *114*, 10235–10253.

[S7] Kollman, P.; *Chem. Rev.* **1993**, *93*, 2395–2417.

APPENDIX B

Python Package: Seconday Structure Builder

B.1 Code description

This python package creates protein coordinates file in the PDB formate for a given amino acid sequence and in desired secondary structure. This code consists of three main scripts: `Anglefilemaker`, `Polypeptidemaker`, and `Sidechainsticker`.

- `Anglefilemaker.py:` This code reads desired structure as an input and creates the list of dihedral angles accordingly.
- `Polypeptidemaker.py:` This code creates polypeptide backbone chain using predefined dipeptide coordinates; then, applies obtained dihedral angle to produce final backbone coordinates.
- `Sidechainsticker.py:` This code reads given sequence as an input and uses predefined sidechain coordinates in order to add sidechain group onto the existing polypeptide backbone and creates final protein structure.

Common functions used by above code is loaded as python module (`ss-builder-module.py`) into the scripts. Graphical user interface, `ProteinMaker.py`, also written by Emilie Fortin to make the package more convenient.

B.2 Anglefilemaker.py

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-
################################################
#                                              #
# Anglefilemaker.py written by Mehran Bagheri,  #
# Edited by Emilie Fortin                      #
# Dr.James Harden Group, University of Ottawa   #
# contact     : mbagh024@uottawa.ca,           #
# contact     : efort063@uottawa.ca,           #
# Last update : Oct 2016                       #
#                                              #
```

```
################################################
################################################
#                                              #
# This program, using the sequence and structure#
# files as input, calculates the dihedral angles#
# (phi and psi) and writes them onto the       #
# dihedralangle output file.                   #
#                                              #
# To Compile :                                 #
# python python/Anglefilemaker.py              #
#                                              #
# The structure file contains the Type of      #
# structure as well as the number of residues in#
# each structure. Each type of structure is    #
# defined in the variables section below       #
#                                              #
# The Sequence file contains the sequence using #
# the proteins sequence using the 1-letter code#
#                                              #
################################################

#--------------------------------------------------------------------#
# VARIABLES:
#--------------------------------------------------------------------#
# phi[] = The angle between the alpha-carbon and the Nitrogen for each amino acid
# psi[] = The angle between the alpha-carbon and the Carbon for  each amino acid
# Type[] = Type of structure ex: alpha-helix, beta-sheet etc.
# length[] = Number of amino acids in each structure type
# seq[] = Sequence of amino acids in the protein in 1 letter code format
# The definition of the Type[] of structures:
# AHr = Alpha-helix right handed
# AHl = Alpha-helix left handed
# 310H = 3-10 helix
# Pi = Pi helix
# BS = Beta strand (without any backbone torsion - completely straight)
# PPl = Polyproline left handed helix
# PPr = Polyproline right handed helix
# BT6 = 6 molecules in turn (no proline) - Creates beta sheet
# BT2 = 2 molecules in turn - Creates beta sheet
# BT30 = Creates a 30 degree turn between two beta strands
# BT60 = Creates a 60 degree turn between two beta strands
# BT90 = Creates a 90 degree turn between two beta strands
# BT120 = Creates a 120 degree turn between two beta strands
# R = Random structure

#--------------------------------------------------------------------#

################################################
#                       Main                   #
################################################
def Anglefm():
  #-----------------------------------------------
  # function & classes
```

```
#-----------------------------------------------

import sys, math, random;
print '\n Anglefilemaker.py : ...\n';

#-----------------------------------------------
# Loading input data
#-----------------------------------------------
Fin1 = 'inputs/structure.txt' ;
Fin2 = 'inputs/sequence.txt' ;
Fout = 'outputs/dihedralangle.txt';

print '\t * loads input data:';
print '\t   ** "%s"'%Fin1;
print '\t   ** "%s"'%Fin2;
# Opens Fin-adress for reading the Structure as input
Structurefile = open(Fin1, "r");
Sequencefile = open(Fin2, "r");
# Makes file in Fout-address with the name of
#Fout argument and write on it as an output
ofile = open(Fout, "w");
structure = Structurefile.readlines();
sequence = Sequencefile.readlines();
Structurefile.close();
Sequencefile.close();

# Makes a list of the Types of secondary
#structure and the length of each structure
# Type of structure alphahelix, beta , ... and the length of them
ind = []; struc = []; number = [];
for line in structure:
  indval, strucval, numberval = line.split();
  # Reads the input data : structure and length
  ind.append(indval); struc.append(strucval); number.append(int(numberval));

seq = [];
for line in sequence:
  seqval, rotangval = line.split();
  seq.append(seqval);
if len(seq) == 0 or len(seq) == 1 :
  exit('\n Error (1): wrong number of residue in "sequence" file (need > 1)\n')

# make list secondary strcuture
ss_list = [];index = [];
for i in range(len(ind)):
  ssval = struc[i];
  for j in range(number[i]):
    ss_list.append(ssval);
    # index of residue in secondary strcuture
    index.append(j+1);

# examine total length of strcuture = number of residue in seguence
```

```
if len(ss_list) != len(seq):
  exit("\n Error(2): number of given strcuture:%s must be ...
  ...  equal to\n lenght of given sequence:%s")%(len(ss_list),len(seq));
print '\n\t * Number of residues (alpha-carbons): %s' %(len(seq));

#-----------------------------------------------
# Calculates the psi & phi angles for each alpha-carbon
#-----------------------------------------------
i=0;
while i < len(ss_list):
  ### Helical Structure ###
  if ss_list[i] == 'AHr':         # ALPHA HELIX right-handed (-60,-45)
    ofile.write('-60\t-45\n'); i=i+1;
  elif ss_list[i] == 'AHl':       # ALPHA HELIX left-handed (60,45)
    ofile.write('60\t45\n'); i=i+1;
  elif ss_list[i] == '310H':      # 3_10 HELIX (-50,-25)
    ofile.write('-50\t-25\n'); i=i+1;
  elif ss_list[i] == 'Pi':        # Pi HELIX (-55,-70)
    ofile.write('-55\t-70\n'); i=i+1;
  elif ss_list[i] == 'PPl':       # Poly-proline left handed helix (-75,150)
    ofile.write('-75\t150\n'); i=i+1;
  elif ss_list[i] == 'PPr':       # Poly-proline right handed helix (-75,160)
    ofile.write('-75\t160\n'); i=i+1;
  ### Beta Strsucture ###
  # completely straight, no torsion along the chain more precisely (-134,130.95)
  elif ss_list[i] == 'BS':
    ofile.write('-134\t131\n'); i=i+1;
  # completely straight beta-strand without any twist (-122,118)
  elif ss_list[i] == 'BS1':
    ofile.write('-122\t118\n'); i=i+1;
  elif ss_list[i] == 'test':        # test
    ofile.write('-120\t135\n'); i=i+1;
  ### Beta turn ###
  elif ss_list[i] == 'BT6':       # BETA SHEET Turn 6 residues in turn
    ofile.write('-130\t120\n-140\t150\n20\t-80\n-120\t-70\n-30\t ...
    ... 150\n-165\t110\n');
    i=i+6;
  # BETA SHEET Turn 6 residues in turn - proline allowed on 2nd position
  elif ss_list[i] == 'BT6P':
    ofile.write('-90\t110\n-72\t-28\n-60\t-61\n-87\t-16\n82\t48\n-113\t130\n');
    i=i+6;
   # BETA SHEET Turn 2 residues in turn (0,90)(90,0)
  elif ss_list[i] == 'BT2':
    ofile.write('0\t90\n90\t0\n'); i=i+2;
  # BETA SHEET Turn 4 residues in turn
  elif ss_list[i] == 'BT4':
    ofile.write('0\t90\n90\t0\n'); i=i+2;
  # for Beta Turn-90 degrees - proline allowed on 1st position
  elif ss_list[i] == 'BT90':
    #(-60,-30)(-120,120)
    ofile.write('-60\t-30\n-120\t120\n'); i=i+2;
  # for Beta Turn-60 degrees - proline allowed on 1st position
```

```
    elif ss_list[i] == 'BT60':
      #(-60,-30)(-90,0) #65deg
      ofile.write('-60\t-30\n-90\t0\n'); i=i+2;
    # for Beta Turn-30 degrees (60,30)(90,0)# 32deg
    elif ss_list[i] == 'BT30':
      ofile.write('60\t30\n90\t0\n'); i=i+2;
    # for Beta Turn-120 degrees - proline allowed on 2nd position
    elif ss_list[i] == 'BT120':
      #(-120,120)(-60,0) #115deg
      ofile.write('-120\t120\n-60\t0\n'); i=i+2;
    ### Random coil ###
    elif ss_list[i] == 'R':
      exit('\nError (MB): Need to be completed\n');
      # random.randint generates a random integer between 0 and 180
      Rleft= 45 + random.randint(0,180);
      if Rleft > 180:
        Rleft = Rleft - 360
      # This helps generate the random phi and psi angles for the sequence
      Rright= 45 - random.randint(0,180);
      if Rright < -180:
        Rright = Rright + 360
      # Writes the output onto the output file
      ofile.write('%s\t%s\n' %(Rleft,Rright)); i=i+1;
    ### Unknown structure ###
    else:
      exit('\nError (3): Unknown structure %s\n' %ss_list[i]);

  ofile.close()

  #-------------------------------------------------
  # Corrects the phi angles for proline and Writes
  # the data in the output file (dihedralangle.txt)
  #-------------------------------------------------
  print '\n\t * WARNING:'
  print '\t   proline phi angle is rescricted to ...
   ... approximately -60, setting proline to'
  print '\t   different angle, may locally ...
   ... destabilize the configured structure'

  ofile = open(Fout, "r")      # Opens the output file to read it
  angle = ofile.readlines();    # Reads the lines
  ofile.close()
  phi=[];psi=[];       # phi is the angle between the alpha-c and the N ...
  for line in angle:    # ... and psi is the angle between the alpha-c and the C
    # phi is the first column and psi the second in the file
    phival, psival = line.split();
    phi.append(phival); psi.append(psival);

  for i in range(len(phi)):
    if seq[i]== 'P' and phi[i] != '-60' :
      print '\t   ** Warning: Pro%s , structure:%s , phi:%s'%(i,ss_list[i],phi[i]);
```

```
  print '\n Anglefilemaker.py : Done\n';

#Anglefm()
##################################################
#                       End                      #
##################################################
## TEMPLATES FOR CREATING CUSTOM ANGLES ##
# Template for a repeated sequence
#    elif ss_list[j] == 'CODE':        # NAME OF STRUCTURE
#      ofile.write(' ');
```

B.3 Polypeptidemaker.py

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-
##################################################
#                                                #
# Polypeptidemaker.py written by Mehran Bagheri #
# Edited by Emilie Fortin                        #
# Dr.James Harden Group, University of Ottawa    #
# contact     : mbagh024@uottawa.ca,             #
# contact     : efort063@uottawa.ca,             #
# Last update : May 2016                         #
#                                                #
##################################################
##################################################
#                                                #
# To Compile :                                   #
# python python/Polypeptidemaker.py              #
#                                                #
# Input data:                                    #
#   * (1)  dihedralangle                         #
#   * (2) peptide-coordination                   #
#                                                #
# dihedralangle:                                 #
#   This file contains the phi and psi angles    #
#   for each alpha-carbon in the chain as        #
#   calculated by anglefilemaker.py              #
#                                                #
# peptide-coordination.pdb:                      #
#   This file contains the structure of a short  #
#   polypeptide containing only 2 amino acids    #
#   with (phi, psi) angles of (180,180) which    #
#   is the smallest unit that can be repeated    #
#   to obtain a polypeptide chain. This          #
#   structure is called the peptide unit cell.   #
#   (The sidechains are omitted from this model  #
```

```
#   and have R as a stand in)                    #
#                                                #
# polypeptide-chain.pdb (output):                #
#   This file contains the coordinates of the    #
#   polypeptide backbone of the desired length   #
#   without any side chains                      #
#                                                #
# How this code works:                           #
#   It obtains the length of the protein from    #
#   the input dihedralangle file, and uses the   #
#   peptide-coordination as a building block to  #
#   make a polypeptide of the desired length.    #
#   Then it applies the angles from the          #
#   dihedralangle file to the chain and exports  #
#   the structure in a pdb file named            #
#   polypeptide-chain.                           #
#                                                #
##################################################

#-----------------------------------------------------------------------#
# VARIABLES:
#-----------------------------------------------------------------------#
# atom_type[]= Atom type for atoms of peptide unit cell
# ox[] = x-coordinate for atoms of peptide unit cell (MB:
#coordinates for atoms of peptide unit cell)
# oy[] = y-coordinate for atoms of peptide unit cell  (MB:
#call variable with more meaningfull name)
# oz[] = z-coordinate for atoms of peptide unit cell
# phi[] = the phi angle between the alpha carbon and the
#nitrogen for an entry [] in the list
# psi[] = the psi angle between the alpha carbon and the
#other carbon for an entry [] in the list
# n = index number
# BblockN = ...
#-----------------------------------------------------------------------#

##################################################
#                      Main                      #
##################################################

def Polypm():
  #-----------------------------------------------
  # function & classes
  #-----------------------------------------------
  import sys, math, random;
  from ss_builder_module import *;
  print '\n Polypeptidemaker.py : ...\n';
  #-----------------------------------------------
  # Loading input data
  #-----------------------------------------------
  # Load coordination of peptide unit-cell and dihedral angles
  Fin1='library/peptide-coordination.pdb';
  Fin2='outputs/dihedralangle.txt';
```

```
Fin3 = 'inputs/sequence.txt';
print '\t * loads input data:'
print '\t   ** "%s"'%Fin1;
print '\t   ** "%s"'%Fin2;
print '\t   ** "%s"'%Fin3;
peptidefile = open(Fin1, "r");
anglefile = open(Fin2, "r");
sequencefile = open(Fin3, "r");
coordination = peptidefile.readlines();
angle = anglefile.readlines();
sequence = sequencefile.readlines();
anglefile.close(); peptidefile.close();sequencefile.close();
if len(angle) == 0 or len(angle) == 1 :
  exit('\n Error (1): wrong number of angle in ...
   ... "dihedralsangle.txt" file (need > 1)\n')

seq = [];
for line in sequence:
  seqval, rotangval = line.split();
  seq.append(seqval);

# Makes a list of original input coordinates
#(Peptide-coordination and dihedralangle)
atom_type = []; ox = []; oy = []; oz = [];   # orginal coordinates
for line in coordination:
  atomval, atom_numval, atom_typeval, res_numval, ...
   ... oxval, oyval, ozval = line.split();
  atom_type.append(atom_typeval);
  ox.append(float(oxval)); oy.append(float(oyval)); oz.append(float(ozval));
phi = []; psi = [];     # phi and psi values for each amino acid
for line in angle:
  leftangleval,rightangleval = line.split();
#Since the building block already has angles of (180,180) for phi,psi
  phi.append(float(leftangleval)-180.); psi.append(float(rightangleval)-180.);
print '\n\t * Desired length of the polypeptide: %s amino-acids' %(len(angle))

#-------------------------------------------------
# Creates polypeptide chain with desire length
#-------------------------------------------------
print '\t * Creates a polypeptide chain from N-terminus to C-terminus'
# The unit block vector represents the distance
# between everyother alpha carbons which is equal to
# position of third alpha carbon:14CA in unite cell corrdinates.
# first CA is position at origine.
unit_block_vec = [ox[13],oy[13],oz[13]]
# Based on the unit cell, for n shift you would have 2n+1
# complete amino acid bakcbone (where complete
# indicates that it has it's own CA, N-H, C-O, and R)
# for even number of residues, an extra residue
# created but wont be written into the output
BblockN = len(angle)/2;
n = 0;x = [];y = [];z = [];
```

```
while n <= BblockN:        # shift unit cell atoms for n times
  for i in range(len(ox)):
    xval = (n*unit_block_vec[0])+ox[i];
    yval = (n*unit_block_vec[1])+oy[i];
    zval = (n*unit_block_vec[2])+oz[i];
    x.append(xval); y.append(yval); z.append(zval);
  n = n+1;

#-----------------------------------------------
# Applying the phi and psi angles to each alpha-carbon in the chain
#-----------------------------------------------
n = 0 ;
while n < len(angle):
# Sets the alpha-carbon as the origin
  o = [0,0,0];
  # Sets the alpha-carbon to be the origin
  o[0] = x[6+(7*n)];o[1] = y[6+(7*n)];o[2] = z[6+(7*n)];
  change_origin(x,y,z,o);
# Rotates around the CA-N bond (applies the phi angle)
  s = 0; v =[0,0,0];       # For the left side of middle alpha carbon
   # alphaC-N is the vector = N-7 when C = 0,0,0
  v[0] = x[3+(7*n)];v[1] = y[3+(7*n)];v[2] = z[3+(7*n)];
  # Takes all the atoms before (to the left of) the
  # alpha-C except it's H and R
  while s <= (3+(7*n)) :
    # and rotates them around the alphaC-N bond by the angle phi
    x[s],y[s],z[s] = rotate_xyz_v(x[s],y[s],z[s],v,phi[n]);
    s = s+1;
# Rotates around the CA-C bond (applies the psi angle)
  s = 7+(7*n); v =[0,0,0]; # For the right side of middle alpha carbon
   # alphaC-C is the vector = C-11 when alphaC = 0,0,0
  v[0] = x[8+(7*n)];v[1] = y[8+(7*n)];v[2] = z[8+(7*n)];
  # Takes all atoms after (to the right of) the
  # alpha-C except it's H and R
  while s < len(x) :
    # and rotates them around the alphaC-C bond by the angle psi
    x[s],y[s],z[s] = rotate_xyz_v(x[s],y[s],z[s],v,psi[n]);
    s = s+1;
  o = [-i for i in o]      # Inverts the origin vector
  # Returns everything back to it's previous location
  # while keeping the new angles
  change_origin(x,y,z,o);
  n = n+1;
# Shift origin to have all positive coordinates
o[0] = min(x);o[1] = min(y);o[2] = min(z);
change_origin(x,y,z,o);

#-----------------------------------------------
# Writes the data in the output file (output.pdb) in pdb format
#-----------------------------------------------
# Creates output file to write backbone coordinates
Fout1='outputs/polypeptide-chain.pdb';
ofile = open(Fout1, "w");
```

```
  print '\t * Writes the coordinates of the ...
  ...  polypeptide chain into:\n\t   ** "%s"'%(Fout1)
  # Each complete alpha carbon has 7 atoms, thus the 7 atom repeat.
  # We need to exclude the first C and O in order
  # for the sequence to start at the N-terminus
  for i in range(2,2+7*len(angle)):
    # The first 3 conditions make sure that the pdb
    # formating for the x,y,z coordinates can be respected
    if x[i] >= 10000:
      exit('\n Error (2): x component of atom %g has ...
      ... more digits than maximum distance \n'%(i));
    elif y[i] >= 10000:
      exit('\n Error (2): y component of atom %g has ...
      ... more digits than maximum distance \n'%(i));
    elif z[i] >= 10000:
      exit('\n Error (2): z component of atom %g has ...
      ... more digits than maximum distance \n'%(i));
    # To remove unnecessary H in the Proline backbone
    if not (atom_type[i%14] == 'H' and seq[((i-2)/7)] == 'P'):
      # Writes the output in PDB format
      ofile.write('ATOM  %5d %-4s     %4d    %8.3f%8.3f%8.3f\n' ...
      ... %(i-1,atom_type[i%14],1+((i-2)/7),x[i],y[i],z[i]));
  # Writes the bond conection for the pdb file
  ofile.write('TER\n')
  ofile.close();
  print '\n Polypeptidemaker.py : Done\n';
#Polypm()
##################################################
#                      End                       #
##################################################
```

B.4 Sidechainsticker.py

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-
##################################################
#                                                #
# sidechainsticker.py written by Mehran Bagheri,#
# Edited by Emilie Fortin                        #
# Dr.James Harden Group, University of Ottawa    #
# contact     : mbagh024@uottawa.ca,             #
# contact     : efort063@uottawa.ca,        #
# Last update : May 2016                         #
#                                                #
##################################################
##################################################
#                                                #
```

```
# This program gives you the final protein      #
# by attaching the side chains onto the existing#
# polypeptide backbone created using            #
# polypeptidemaker.py                           #
#                                               #
# To Compile :                                  #
# python/Sidechainsticker-Final.py              #
#                                               #
# side-chain: Contains the coordinates for each #
#             of the sidechains  [input file]   #
#                                               #
# Sequence: This file contains the sequence of  #
#           amino acids in 1 letter code along  #
#           with the angles of the side chains  #
#           (Theta) [input file]                #
#                                               #
# polypeptide-chain: Contains the coordinates   #
#                   for the polypeptide backbone#
#                   [input file]                #
#                                               #
# output: This is the final output which gives  #
#         The coordinates of the complete       #
#         polypeptide in PDB format to be read  #
#         by VMD                                #
#                                               #
#################################################

#-----------------------------------------------------------------#
# VARIABLES:
#-----------------------------------------------------------------#
#
# AA, RA, NA, DA, CA, QA, EA, GA, HA, IA, LA, KA, MA, FA, PA, SA, TA, WA, YA, VA
#= The Name of each atom in each side chain divided up by amino acid (the first
#letter is the 1 letter code)
# Ax, Rx, Nx, Dx, Cx, Qx, Ex, Gx, Hx, Ix, Lx, Kx, Mx, Fx, Px, Sx, Tx, Wx, Yx, Vx
#= The x coordinates of each atom in the side-chain file divided up by amino acid
#(the first letter is the 1 letter code)
# Ay, Ry, Ny, Dy, Cy, Qy, Ey, Gy, Hy, Iy, Ly, Ky, My, Fy, Py, Sy, Ty, Wy, Yy, Vy
#= The y coordinates of each atom in the side-chain file divided up by amino acid
#(the first letter is the 1 letter code)
# Az, Rz, Nz, Dz, Cz, Qz, Ez, Gz, Hz, Iz, Lz, Kz, Mz, Fz, Pz, Sz, Tz, Wz, Yz, Vz
#= The z coordinates of each atom in the side-chain file divided up by amino acid
#(the first letter is the 1 letter code)
# x = The x coordinates of each atom in the side-chain file
# y = The y coordinates of each atom in the side-chain file
# z = The z coordinates of each atom in the side-chain file
# number = Atom serial number for each atom in the side-chain file
# group = 3 letter code name of amino acid
# Type = Name of each atom in the sidechain
# Seq =  1 letter code for each amino acid in the sequence file
# Theta = Angle of the side chains - zero by default
#*(will later be rotated to prevent steric hinderance)*
# TYPE = Name of each atom in the backbone (eg: CA, HA, H, N, O, C)
```

```
# GROUPN = Residue sequence number for each atom in the polypeptide-chain file
# NUMBER = Atom serial number for each atom in the polypeptide-chain file
# X = The x coordinates of each atom in the polypeptide-chain file
# Y = The y coordinates of each atom in the polypeptide-chain file
# Z = The z coordinates of each atom in the polypeptide-chain file
#
#--------------------------------------------------------------------#

##################################################
#                     Main                       #
##################################################

#------------------------------------------------
# function & classes
#------------------------------------------------

import sys, math, random;
from ss_builder_module import *;


#------------------------------------------------
# Loading input data
#------------------------------------------------
def Sidecs(Fout):
  print '\n Sidechainsticker.py : ...\n'
  Fin1 = 'library/side-chain.txt';
  Fin2 = 'inputs/sequence.txt';
  Fin3 = 'outputs/polypeptide-chain.pdb';
  print '\t * loads input data:'
  print '\t   ** "%s"'%Fin1;
  print '\t   ** "%s"'%Fin2;
  print '\t   ** "%s"'%Fin3;

  # Opens F1in-adress for reading the aminoacid coordinations as input
  sidechainfile = open(Fin1, "r");
  # Opens F2in-adress for reading the aminoacid coordinations as input
  sequencefile = open(Fin2, "r");
  # Opens F3in-adress for reading the aminoacid coordinations as input
  proteinchainfile = open(Fin3, "r");
  sidechain = sidechainfile.readlines();
  sidechainfile.close();
  sequence = sequencefile.readlines();
  sequencefile.close();
  proteinchain = proteinchainfile.readlines();
  proteinchainfile.close();


  # Extracting information from the side-chain file
  # A list of all the names of the atoms in each amino acid group
  AA=[];RA=[];NA=[];DA=[];CA=[];QA=[];EA=[];GA=[];HA=[];IA=[];
  LA=[];KA=[];MA=[];FA=[];PA=[];SA=[];TA=[];WA=[];YA=[];VA=[];
  # A list of all the x coordinates of each atom in a given amino acid
```

```
Ax=[];Rx=[];Nx=[];Dx=[];Cx=[];Qx=[];Ex=[];Gx=[];Hx=[];Ix=[];
Lx=[];Kx=[];Mx=[];Fx=[];Px=[];Sx=[];Tx=[];Wx=[];Yx=[];Vx=[];
# A list of all the y coordinates of each atom in a given amino acids
Ay=[];Ry=[];Ny=[];Dy=[];Cy=[];Qy=[];Ey=[];Gy=[];Hy=[];Iy=[];
Ly=[];Ky=[];My=[];Fy=[];Py=[];Sy=[];Ty=[];Wy=[];Yy=[];Vy=[];
# A list of all the z coordinates of each atom in a given amino acid
Az=[];Rz=[];Nz=[];Dz=[];Cz=[];Qz=[];Ez=[];Gz=[];Hz=[];Iz=[];
Lz=[];Kz=[];Mz=[];Fz=[];Pz=[];Sz=[];Tz=[];Wz=[];Yz=[];Vz=[];
x=[];y=[];z=[];number=[];group=[];Type=[];
# This loop ensures that every line in the document gets copied
for line in sidechain:
  # Copies the values of each variable from the side chain file
  groupval,numberval,Typeval,xval,yval,zval = line.split();
  group.append(groupval);number.append(int(numberval));Type.append(Typeval)
  x.append(float(xval));y.append(float(yval));z.append(float(zval));
# Puts each parameter in its own category
# For instance puts all the atoms in the alanine sidechain into A
for s in number:
  # Put coordinates for each specific atom in Ax,Ay,Ay as well as the atom's name
  if group[s-1] == 'ALA':
    AA.append(Type[s-1]);Ax.append(x[s-1]);Ay.append(y[s-1]);Az.append(z[s-1]);
  elif group[s-1] == 'ARG':
    RA.append(Type[s-1]);Rx.append(x[s-1]);Ry.append(y[s-1]);Rz.append(z[s-1]);
  elif group[s-1] == 'ASN':
    NA.append(Type[s-1]);Nx.append(x[s-1]);Ny.append(y[s-1]);Nz.append(z[s-1]);
  elif group[s-1] == 'ASP':
    DA.append(Type[s-1]);Dx.append(x[s-1]);Dy.append(y[s-1]);Dz.append(z[s-1]);
  elif group[s-1] == 'CYS':
    CA.append(Type[s-1]);Cx.append(x[s-1]);Cy.append(y[s-1]);Cz.append(z[s-1]);
  elif group[s-1] == 'GLN':
    QA.append(Type[s-1]);Qx.append(x[s-1]);Qy.append(y[s-1]);Qz.append(z[s-1]);
  elif group[s-1] == 'GLU':
    EA.append(Type[s-1]);Ex.append(x[s-1]);Ey.append(y[s-1]);Ez.append(z[s-1]);
  elif group[s-1] == 'GLY':
    GA.append(Type[s-1]);Gx.append(x[s-1]);Gy.append(y[s-1]);Gz.append(z[s-1]);
  elif group[s-1] == 'HIS':
    HA.append(Type[s-1]);Hx.append(x[s-1]);Hy.append(y[s-1]);Hz.append(z[s-1]);
  elif group[s-1] == 'ILE':
    IA.append(Type[s-1]);Ix.append(x[s-1]);Iy.append(y[s-1]);Iz.append(z[s-1]);
  elif group[s-1] == 'LEU':
    LA.append(Type[s-1]);Lx.append(x[s-1]);Ly.append(y[s-1]);Lz.append(z[s-1]);
  elif group[s-1] == 'LYS':
    KA.append(Type[s-1]);Kx.append(x[s-1]);Ky.append(y[s-1]);Kz.append(z[s-1]);
  elif group[s-1] == 'MET':
    MA.append(Type[s-1]);Mx.append(x[s-1]);My.append(y[s-1]);Mz.append(z[s-1]);
  elif group[s-1] == 'PHE':
    FA.append(Type[s-1]);Fx.append(x[s-1]);Fy.append(y[s-1]);Fz.append(z[s-1]);
  elif group[s-1] == 'PRO':
    PA.append(Type[s-1]);Px.append(x[s-1]);Py.append(y[s-1]);Pz.append(z[s-1]);
  elif group[s-1] == 'SER':
    SA.append(Type[s-1]);Sx.append(x[s-1]);Sy.append(y[s-1]);Sz.append(z[s-1]);
  elif group[s-1] == 'THR':
    TA.append(Type[s-1]);Tx.append(x[s-1]);Ty.append(y[s-1]);Tz.append(z[s-1]);
```

```
  elif group[s-1] == 'TRP':
    WA.append(Type[s-1]);Wx.append(x[s-1]);Wy.append(y[s-1]);Wz.append(z[s-1]);
  elif group[s-1] == 'TYR':
    YA.append(Type[s-1]);Yx.append(x[s-1]);Yy.append(y[s-1]);Yz.append(z[s-1]);
  elif group[s-1] == 'VAL':
    VA.append(Type[s-1]);Vx.append(x[s-1]);Vy.append(y[s-1]);Vz.append(z[s-1]);
  else : sys.exit('\n\tError (1): Undefined aminoacid in input file: "%s"'%Fin1);

# Several lists information about every amino acid to
#simplify the future steps
# AAlist3: includes the 3 letter code for each
#amino acid in alphabetical order
# AAlist1: includes the single letter code for each
# amino acid in same order as AAlist3
# AAlength: gives the length (number of atoms) of
#each amino acid sidechain in the same order
AAlist3 =['ALA','ARG','ASN','ASP','CYS','GLU','GLN','GLY','HIS',
   ...'ILE','LEU','LYS','MET','PHE','PRO','SER','THR','TRP','TYR','VAL']
AAlist1 = ['A','R','N','D','C','Q','E','G','H', ...
... 'I','L','K','M','F','P','S','T','W','Y','V']
AAlength = [5,19,9,7,6,11,12,2,12,14,14,17,11,15,10,6,9,19,16,11]

# Extracting the information from the sequence file
# The Sequence is the 1 letter code for each amino acid in
# the polypeptide sequence and Theta is the sidechain angle
Seq=[];Theta=[];
for line in sequence:
  Seqval, Thetaval = line.split();
  Seq.append(AAlist3[AAlist1.index(Seqval)]); Theta.append(float(Thetaval));
# Extracting the information from the proteinchain file
TYPE=[];GROUPN=[];NUMBER=[];X=[];Y=[];Z=[];
for line in proteinchain:
  if line[0:4] == 'ATOM':       # Only reads atom coordinate line
     # Copies the values of each variable from the proteinchain file
    ATOMval,NUMBERval,TYPEval,GROUPNval,Xval,Yval,Zval=line.split();
    TYPE.append(TYPEval);GROUPN.append(int(GROUPNval));
    X.append(float(Xval));Y.append(float(Yval));Z.append(float(Zval));
print '\t * Number of amino acids in sequence file: %s' %(len(Seq));
print '\t * Number of alpha-C in polypeptide chain: %s' %(int(GROUPNval));
if int(GROUPNval) != len(Seq) :
  sys.exit('\n Error (2): sequence file is not equal to polepeptide chain');

#-------------------------------------------------
# Attaches the side chains to the polypeptide backbone
#-------------------------------------------------
i=0;j=0;FCX=[];FCY=[];FCZ=[];   # j counter for each alpha-C
FType=[];FGroupN=[];FGroup=[];  # Final Type, Final Coordinates
# Following loop goes through the proteinchain and adds ...
# ... the appropriate side chain whenever it reaches an alpha-Carbon
while i < len(X):
  # write associated X,Y,Z coordinates and the type, sequence and
  # group number of atoms H, N, or HA which corrosponds to residue j,
```

```
if (TYPE[i] == 'H' or TYPE[i] =='N' or TYPE[i] =='HA') :
  FCX.append(X[i]);FCY.append(Y[i]);
  FCZ.append(Z[i]);FType.append(TYPE[i]);FGroup.append(Seq[j]);
  FGroupN.append(GROUPN[i]);
elif (TYPE[i] =='C' or TYPE[i] == 'O') :
  FCX.append(X[i]);FCY.append(Y[i]);
  FCZ.append(Z[i]);FType.append(TYPE[i]);FGroup.append(Seq[j-1]);
  FGroupN.append(GROUPN[i]);
elif TYPE[i]=='CA' :
  # Defines a vector V that will be the side chain vector.
  V = [X[i-2]-X[i],Y[i-2]-Y[i],Z[i-2]-Z[i]];
  if Seq[j] not in AAlist3:
    sys.exit('\n\tError (4): Undefined aminoacid in input sequence file\n')
  else:
    # Makes a copy of sidechain coordinates
    s=0;samplex=[];sampley=[];samplez=[];
    w = AAlist3.index(Seq[j])                # get the index of  seq[j]
    while s < AAlength[w]:
      # copy the corresponding x, y, z coordinates
      # for the side chain of that amino acid
      if w==0: samplex.append(Ax[s]); sampley.append(Ay[s]); ...
      ... samplez.append(Az[s]);s=s+1;   # ALA
      elif w==1: samplex.append(Rx[s]); sampley.append(Ry[s]); ...
       ... samplez.append(Rz[s]);s=s+1; # ARG
      elif w==2: samplex.append(Nx[s]); sampley.append(Ny[s]); ...
       ... samplez.append(Nz[s]);s=s+1; # ASN
      elif w==3: samplex.append(Dx[s]); sampley.append(Dy[s]); ...
       ... samplez.append(Dz[s]);s=s+1; # ASP
      elif w==4: samplex.append(Cx[s]); sampley.append(Cy[s]); ...
       ... samplez.append(Cz[s]);s=s+1; # CYS
      elif w==5: samplex.append(Ex[s]); sampley.append(Ey[s]); ...
       ... samplez.append(Ez[s]);s=s+1; # GLU
      elif w==6: samplex.append(Qx[s]); sampley.append(Qy[s]); ...
       ... samplez.append(Qz[s]);s=s+1; # GLN
      elif w==7: samplex.append(Gx[s]); sampley.append(Gy[s]); ...
       ... samplez.append(Gz[s]);s=s+1; # GLY
      elif w==8: samplex.append(Hx[s]); sampley.append(Hy[s]); ...
       ... samplez.append(Hz[s]);s=s+1; # HIS
      elif w==9: samplex.append(Ix[s]); sampley.append(Iy[s]); ...
       ... samplez.append(Iz[s]);s=s+1; # ILE
      elif w==10: samplex.append(Lx[s]); sampley.append(Ly[s]); ...
       ... samplez.append(Lz[s]);s=s+1;# LEU
      elif w==11: samplex.append(Kx[s]); sampley.append(Ky[s]); ...
       ... samplez.append(Kz[s]);s=s+1;# LYS
      elif w==12: samplex.append(Mx[s]); sampley.append(My[s]); ...
       ... samplez.append(Mz[s]);s=s+1;# MET
      elif w==13: samplex.append(Fx[s]); sampley.append(Fy[s]); ...
       ... samplez.append(Fz[s]);s=s+1;# PHE
      elif w==14: samplex.append(Px[s]); sampley.append(Py[s]); ...
       ... samplez.append(Pz[s]);s=s+1;# PRO
      elif w==15: samplex.append(Sx[s]); sampley.append(Sy[s]); ...
       ... samplez.append(Sz[s]);s=s+1;# SER
      elif w==16: samplex.append(Tx[s]); sampley.append(Ty[s]); ...
```

```
        ... samplez.append(Tz[s]);s=s+1;# THR
      elif w==17: samplex.append(Wx[s]); sampley.append(Wy[s]); ...
        ... samplez.append(Wz[s]);s=s+1;# TRP
      elif w==18: samplex.append(Yx[s]); sampley.append(Yy[s]); ...
        ... samplez.append(Yz[s]);s=s+1;# TYR
      else : samplex.append(Vx[s]); sampley.append(Vy[s]); ...
        ... samplez.append(Vz[s]);s=s+1;     # VAL
    # This matches sidechain vector to the vector V
    match_vector_3D(samplex,sampley,samplez,V);
    # Then the x,y,z coordinates are rotated around
    # the vector which is now an axis by Theta
    rotate_vectorxyz_list(samplex,sampley,samplez,V,Theta[j]);
    # A new vector v is defined which is the
    # position of the alpha-carbon on the backbone
    v = [X[i],Y[i],Z[i]];
    # Then the x,y,z coordinates of the side chain are
    # shifted to cooincide with that position
    shift_v_list(samplex,sampley,samplez,v);
    s=0;
    # Ensures that the entire length of the sidechain is copied
    while s < AAlength[w] :
      FCX.append(samplex[s]);FCY.append(sampley[s]);
      FGroupN.append(GROUPN[i]);
      # Adds the x,y,z coordinates of the main
      # chain and group number values to the list
      FCZ.append(samplez[s]);
      #The following section adds the names of the atoms in the
      # side chains depending on which amino acid it is
      # as determined by w - the index in AAlist3
      if w==0: FType.append(AA[s]);    # ALA
      elif w==1: FType.append(RA[s]);  # ARG
      elif w==2: FType.append(NA[s]);  # ASN
      elif w==3: FType.append(DA[s]);  # ASP
      elif w==4: FType.append(CA[s]);  # CYS
      elif w==5: FType.append(EA[s]);  # GLU
      elif w==6: FType.append(QA[s]);  # GLN
      elif w==7: FType.append(GA[s]);  # GLY
      elif w==8: FType.append(HA[s]);  # HIS
      elif w==9: FType.append(IA[s]);  # ILE
      elif w==10: FType.append(LA[s]); # LEU
      elif w==11: FType.append(KA[s]); # LYS
      elif w==12: FType.append(MA[s]); # MET
      elif w==13: FType.append(FA[s]); # PHE
      elif w==14: FType.append(PA[s]); # PRO
      elif w==15: FType.append(SA[s]); # SER
      elif w==16: FType.append(TA[s]); # THR
      elif w==17: FType.append(WA[s]); # TRP
      elif w==18: FType.append(YA[s]); # TYR
      else :FType.append(VA[s]);       # VAL
      FGroup.append(Seq[j]) # Adds the sequence number to the list
      s=s+1;
    j=j+1;
i=i+1
```

```
#-----------------------------------------------
# Writes the data in the output file (polypeptide-chain.pdb) in pdb format
#-----------------------------------------------

# Creates the output file to write the coordinates of the final polypeptide
   # Makes file in Fout-address and writes on it as an output
  ofile = open(Fout, "w");
  print '\t * Writes the coordinates of ...
  ... the completed polypeptide into:\n\t   **"%s"'%(Fout)

  n=0;
  # This ensures that the full length of the polypeptide is writen as output
  while n < len(FCX):
# The first condition checks that there aren't too
#  many atoms for the pdb file to handle
    if n>=100000: sys.exit('\n\tError (5): Too many atoms - more than 100000');
# The next three conditions ensure that the X, Y and Z
# components aren't too big to satisfy the formating of the pdb file
    elif FCX[n] <= -1000 or FCX[n] >= 10000: sys.exit('\n\t Error (6): ...
     ... X component of atom %g has more digits than maximum distance\n'%(n));
    elif FCY[n] <= -1000 or FCY[n] >= 10000: sys.exit('\n\t Error (6): ...
     ... Y component of atom %g has more digits than maximum distance\n'%(n));
    elif FCZ[n] <= -1000 or FCZ[n] >= 10000: sys.exit('\n\t Error (6): ...
     ... Z component of atom %g has more digits than maximum distance\n'%(n));
# Writes the coordinates of the final polypeptide in pdb format
# in the output file in order for it to be readable in VMD
    else : ofile.write('ATOM  %5d %-4s %3s %4d    %8.3f%8.3f%8.3f\n' ...
     ... %(n+1,FType[n],FGroup[n],FGroupN[n],FCX[n],FCY[n],FCZ[n]))
    n=n+1;
  ofile.write('TER\n')        # Writes the bond link for the pdb file
  ofile.close();
  print '\n Sidechainsticker.py : Done\n'
#Sidecs('outputs/output.pdb')
################################################
#                      End                     #
################################################
```

B.5 ss-builder-module.py

```
#!/usr/bin/env python

import sys, math, random;

################################################
#           function & classes                 #
################################################
```

```
# Gives the distance between two points
def distance(x1,y1,z1,x2,y2,z2):
  dist = math.sqrt((x1-x2)**2 + (y1-y2)**2 + (z1-z2)**2 );
  return dist;

# Takes three points and gives the angle between them,
def cos_invers(x1,y1,z1,x2,y2,z2,x3,y3,z3):
  f = distance(x1,y1,z1,x2,y2,z2) * distance(x2,y2,z2,x3,y3,z3);
  g = (((x2-x1)*(x2-x3)) + ((y2-y1)*(y2-y3)) + ((z2-z1)*(z2-z3)));
  m = math.acos(g/f);
  m = math.degrees(m);
  return m;

# Makes a unit vector |v| = 1
def unit(v):
  f = math.sqrt((float(v[0]))**2 + (float(v[1]))**2 + (float(v[2]))**2 );
  v[0] = float(v[0])/f; v[1] = float(v[1])/f; v[2] = float(v[2])/f;
  return v;

# Returns all coordinates with respect to the point o
def change_origin(x,y,z,o):
  for i in range(len(x)):
    x[i] = x[i]-o[0]; y[i] = y[i]-o[1]; z[i] = z[i]-o[2];
  return x, y, z;

# Returns central point
def centeral_point(x,y,z):
  o=[0,0,0];SumX=0;SumY=0;SumZ=0;
  for i in range(len(x)):
    SumX = x[i] + SumX; SumY = y[i]+SumY; SumZ = z[i]+SumZ;
  o[0] = float(SumX/len(x));o[1] = float(SumY/len(x));o[2] = float(SumZ/len(x));
  return o;

# Shifts points by vector "v"
def shift_v(x,y,z,v):
    x = x + v[0];
    y = y + v[1];
    z = z + v[2];
    return x, y, z;
# shift allpoints by vector v  (attention :it changes the input
# coordinates because it gets pointer)
def shift_v_list(x,y,z,v):
    s=0;
    while s < len(x):
        x[s] = x[s] + v[0];y[s] = y[s] + v[1];z[s] = z[s] + v[2];
        s=s+1;
    return x, y, z;

# Shifts all points by vector "v"
def shiftall_v(x,y,z,v):
  for i in range(len(x)):
    x[i] = x[i] + v[0];
```

```
    y[i] = y[i] + v[1];
    z[i] = z[i] + v[2];
  return x, y, z;

# Rotates x and y on the x-y plane with respect to the origin
def rotate_origin(x,y,z,angle,atom_number):
  t = math.radians(float(angle))
  C = math.cos(t); S = math.sin(t)
  for s in atom_number:
    X = x[s]*C + y[s]*S ;
    Y = x[s]*(-1.*S) + y[s]*C ;
    x[s] = X; y[s] = Y;
  return x,y,z;

# rotate given point x,y,x around v for give angle
def rotate_xyz_v(x,y,z,v,angle):
  t = math.radians(angle); v = unit(v);
  C = math.cos(t); S = math.sin(t);
  X = ( (C+((v[0]**2)*(1.-C)))*x + ((v[0]*v[1]*(1.-C)) - (v[2]*S))*y + ...
  ...  ((v[0]*v[2]*(1.-C))+(v[1]*S))*z );
  Y = ( ((v[0]*v[1]*(1.-C))+(v[2]*S))*x + (C+((v[1]**2)*(1.-C)))*y + ...
  ...  ((v[1]*v[2]*(1.-C))-(v[0]*S))*z );
  Z = ( ((v[0]*v[2]*(1.-C))-(v[1]*S))*x + ((v[1]*v[2]*(1.-C)) + (v[0]*S))*y + ...
  ...  (C+((v[2]**2)*(1.-C)))*z   );
  x = X; y = Y; z = Z;
  return x,y,z;
# rotate the x,y,x around v (attention :it changes the input
# coordinates because it gets pointer)
def rotate_vectorxyz_list(x,y,z,v,angle):
    t = math.radians(angle); v = unit(v);
    C = math.cos(t); S = math.sin(t); s=0;
    while s < len(x):
        X = ( (C+(v[0]**2)*(1.-C))*x[s]  +  ((v[0]*v[1]*(1.-C)) - v[2]*S)*y[s] ...
        ...  +  ((v[0]*v[2]*(1.-C)) + (v[1]*S))*z[s] );
        Y = ( ((v[0]*v[1]*(1.-C)) + v[2]*S)*x[s]  +  (C+(v[1]**2)*(1.-C))*y[s]...
        ...  +  ((v[1]*v[2]*(1.-C)) - (v[0]*S))*z[s] );
        Z = ( ((v[0]*v[2]*(1.-C)) - v[1]*S)*x[s]  +  ...
        ... ((v[1]*v[2]*(1.-C)) + v[0]*S)*y[s]  +  (C+(v[2]**2)*(1.-C))*z[s]   );
        x[s] = X; y[s] = Y; z[s] = Z;
s = s+1;
    return x,y,z;

# Takes the cross product of A and B to obtain the normal vector
def cross_product(A,B):
  N = [(A[1]*B[2])-(A[2]*B[1]),(A[2]*B[0])-(A[0]*B[2]),(A[0]*B[1])-(A[1]*B[0])];
  return N;

# Given two vectors, it will give you the angle between them
def cross_angle(A,B):
  AB = float(A[0]*B[0] + A[1]*B[1] + A[2]*B[2]);
  RA = ((float(A[0]))**2 + (float(A[1]))**2 + (float(A[2]))**2 )**0.5;
  RB = ((float(B[0]))**2 + (float(B[1]))**2 + (float(B[2]))**2 )**0.5;
  R = AB/(RA*RB);
```

```
  if R <= -1 :
    angle = 180;
  elif R >= 1 :
    angle = 0;
  else:
    angle = math.degrees(math.acos((AB/(RA*RB))));
  return angle;

# It takes the two first components and rotates them to match
# the orientation of the vector V def match_vector_3D(x,y,z,V):
  A=[x[1]-x[0],y[1]-y[0],z[1]-z[0]];N=[];
  N = cross_product(A,V);
  angle = cross_angle(V,A);
  # When the Normal vector is zero, they can only have an angle
  # of 0 or 180 degrees
  if not N == [0,0,0]:
    rotate_vectorxyz_list(x,y,z,N,angle);
  elif angle == 180:         # If the angle is 180 degrees
    for i in range(len(x)):
      x[i]=-x[i];y[i]=-y[i];z[i]=-z[i];
  return x,y,z;
```

B.6 ProteinMaker.py

```
#!/usr/bin/env python
#################################################
#                                               #
# ProteinMaker.py written by Emilie Fortin      #
# Dr.James Harden Group, University of Ottawa   #
# contact     : efort063@uottawa.ca,        #
# Last update : June 2016                       #
#                                               #
#################################################
#################################################
#                                               #
# This code is written in Tkinter to be the GUI #
# interface, it takes the user input and        #
# creates the two input files:                  #
#   * Structure                                 #
#   * Sequence                                  #
#                                               #
# and then it runs the other three programs:    #
#   * Anglefilemaker.py                         #
#   * Polypeptidemaker.py                       #
#   * Sidechainsticker.py                       #
#                                               #
# in order to create the final output file whose#
```

```
# name and location can be selected by the user #
#                                               #
# To run:     python Proteinmaker.py            #
#                                               #
#################################################

#-----------------------------------------------------------------#
# VARIABLES:
#-----------------------------------------------------------------#

################################################
#                    Main                      #
################################################

#----------------------------------------------
# function & classes
#----------------------------------------------

import sys
# Imports the Anglefilemaker program that will be run
from Anglefilemaker import *
# Imports the Polypeptidemaker program that will be run
from Polypeptidemaker import *
# Imports the sidechainsticker program that will be run
from Sidechainsticker import *
try:                          # Pyhton 3
    from tkinter import *
    from tkinter.scrolledtext import ScrolledText
    import filedialog
except ImportError:           # Pyhton 2
    from Tkinter import *
    from ScrolledText import ScrolledText
    import tkFileDialog
    import webbrowser

#----------------------------------------------
# GUI interface
#----------------------------------------------

root = Tk() # Creates the main window
root.grid() # Positions the main window
root.title('Polypeptide maker') # Sets the title of the main window
# Creates label with instructions above Sequence text box
Text1 = Label(root, text='Copy and paste sequence in FASTA format').grid(row=0)
# Creates text box to input sequence
Sequence = ScrolledText(root, width=50, height=10)
Sequence.grid(row=1, columnspan=3)
# Creates label with instructions above Structure text box
Text2 = Label(root, text='Use dropdown menu to select ...
... structure and use entry box to specify number of amino')
Text2.grid(row=2, columnspan=3)
# Creates label with instructions above Structure text box
Text3 = Label(root, text='acids (see help [?] button for more details)')
```

```
Text3.grid(row=3, columnspan=3, sticky=W)
# Creates text box to input Structure
Structure = ScrolledText(root, width=20,height=10)
Structure.grid(row=4, columnspan=1,sticky=W,padx=20)
# Creates text box which will house the sequence associated
#  with each structure type
Info = ScrolledText(root, width=15, height=10)
Info.grid(row=4, column=1, columnspan=2,sticky=W, padx=4)
# A list of all structures in the dropdown menu
ListStructure=['Right Handed Alpha Helix', 'Left Handed ...
... Alpha Helix', '3-10 Helix', 'Pi Helix', 'Beta Strand', ...
... 'Left Handed Polyproline Helix',...
... 'Right Handed Polyproline Helix', 'Beta Sheet Turn (6 residues)',...
... 'Beta Sheet Turn (6 residues - Proline as second amino acid)', ...
... 'Beta Sheet Turn (2 residues)', 'Beta Turn (30 degree turn - 2 residues)', ...
 ... 'Beta Turn(60 degree turn - 2 residues)', ...
... 'Beta Turn (90 degree turn - 2 residues)', ...
 ... 'Beta Turn (120 degree turn - 2 residues)', 'Random Structure']
var1 = StringVar()
# Sets the default structure as the first item in the list
var1.set(ListStructure[0])
# Creates a dropdown menu with all the structure types listed above
S1 = OptionMenu(root, var1, *ListStructure)
S1.grid(row=5, sticky=E)
# Creates an entry window to enter number of
# residues in the sequence selected in the dropdown menu
E1=Entry(root)
E1.grid(row=5, column=1)
# Defines the function that the button Add Structure button will execute
def print_stdout():
  # The dictionnary associates each structures name
  # with a 1 to 4 letter name that the code will later use
  dict='Right Handed Alpha Helix': 'AHr' ,
     'Left Handed Alpha Helix': 'AHl', '3-10 Helix': '310H', 'Pi Helix': 'Pi',
     'Beta Strand': 'BS', 'Left Handed Polyproline Helix': 'PPl',
     'Right Handed Polyproline Helix': 'PPr',
     'Beta Sheet Turn (6 residues)': 'BT6',
     'Beta Sheet Turn (6 residues - Proline as second amino acid)': 'BT6P',
     'Beta Sheet Turn (2 residues)': 'BT2',
     'Beta Turn (30 degree turn - 2 residues)': 'BT30',
     'Beta Turn(60 degree turn - 2 residues)': 'BT60',
     'Beta Turn (90 degree turn - 2 residues)': 'BT90',
     'Beta Turn (120 degree turn - 2 residues)' : 'BT120', 'Random Structure': 'R'
  input1=var1.get() # Takes selected value from the dropdown menu
  input2=E1.get() # Takes number entered into the entry window
  # Defines the index as the number of lines in the Structure textbox
  i=int(Structure.index('end-1c').split('.')[0])
  # Prints the information gathered in the correct format
  Structure.insert(INSERT, '%s\t%s\t%s\n' %(i, dict[input1], input2))
  input3=Sequence.get('1.0', 'end-1c')
  global tot
  if i==1:
    tot=0
```

```
  w=int(E1.get())
  Info.insert(INSERT, '%s%s\t%s%s\n' %(tot+1,input3[tot], ...
  ... input3[tot+w-1],tot+w))
  tot=tot+w
# Creates the "Add structure" button
Addstructure = Button(root, text='Add Structure', command=print_stdout)
Addstructure.grid(row=5, column=2)
# Creates label with instructions about defining output location
Text6 = Label(root, text='Select output location for PDB file')
Text6.grid(row=6, sticky=W)
# Creates entry box to input user selected output file location
Output= Entry(root)
Output.grid(row=6, column=1)
# Defines the function that lets you browse to
# chose the output location and name
def Lookup():
  # Opens a window that will
  name=tkFileDialog.asksaveasfilename(defaultextension='.pdb')
  # Prints the path to the Output box for future use
  Output.insert(INSERT, '%s' %(name))
# Creates button that will allow the user to browse
#  their computer to select output folder
Browse=Button(root, text='Browse', command=Lookup)
Browse.grid(row=6, column=2, sticky=W)
# Defines the function that the clear button will execute
def clear__():
  Structure.delete('1.0', 'end')
  Sequence.delete('1.0','end')
  Info.delete('1.0', 'end')
  E1.delete('0', 'end')
# Creates a button that will clear everything that was input
Clear=Button(root, text='Clear', command=clear__)
Clear.grid(row=7, padx=5, sticky=W)
# Defines the function that the Run button will execute
def __Run__():
 # Gives the adress of the first output file (the sequence file)
  Fout1 = 'inputs/sequence';
  ofile1 = open(Fout1, "w"); # Opens the sequence file
  # Copies the information found in the Sequence textbox
  Seq=Sequence.get("1.0", 'end-1c')
  i=0
  # While loop ensures that each amino acid is
  # copied into the seequence file in it's own line
  while i < len(Seq):
    # Prints the sequence file with the rotation
    # angle of the side chain, zero by default
    ofile1.write('%s\t\t0\n'%(Seq[i]))
    i=i+1
  ofile1.close() # Closes the sequence file
  # Gives the adress of the second output file (the structure file)
  Fout2 = 'inputs/structure';
  ofile2 = open(Fout2, "w"); # Opens the structure file
  # Copies the information found in the Structure textbox
```

```
  St=Structure.get("1.0", 'end-1c')
  ofile2.write(St) # Writes that information onto the Structure file
  ofile2.close() # Closes the structure file
  Anglefm() # Runs the Anglefilemaker code
  Polypm() # Runs the Polypeptidemaker code
  # Gets the final output location for Sidechain sticker to use
  oname = str(Output.get())
  Sidecs(oname)# Runs the Sidechainsticker code
# Creates Run button
Run = Button(root, text='Run', command=__Run__)
Run.grid(row=7, sticky=E)
# Defines the function that the help button will execute
def Help_():
 # The link that will open up when someone clicks on
 # the help button(google until we have our own website)
  url='http://www.google.com'
  webbrowser.open_new(url) # Makes the link open when the function is called
# Creates Help button
Help = Button(root, text='?', command=Help_)
Help.grid(row=7, column=2, ipadx=5, pady=8)

root.mainloop()
##################################################
#                      End                       #
##################################################
```

B.7 Predefined dipeptide coordinates

```
ATOM      1  O           1    1.06666  2.14741  0.0000
ATOM      2  C           1    1.20590  0.92532  0.0000
ATOM      3  H           2    2.45599 -0.64630  0.0000
ATOM      4  N           2    2.39616  0.33187  0.0000
ATOM      5  R           2    3.63249  1.96262  1.2442
ATOM      6  HA          2    3.63249  1.71936 -0.8896
ATOM      7  CA          2    3.63249  1.08950  0.0000
ATOM      8  O           2    4.69915 -1.05791  0.0000
ATOM      9  C           2    4.83839  0.16418  0.0000
ATOM     10  H           3    6.08848  1.73579  0.0000
ATOM     11  N           3    6.02865  0.75762  0.0000
ATOM     12  R           3    7.26498 -0.87312 -1.2442
ATOM     13  HA          3    7.26498 -0.62986  0.8896
ATOM     14  CA          3    7.26498  0.00000  0.0000
```

B.8 Predefined sidechain group coordinates

```
ALA 1 CA 0.000 0.000 0.000
ALA 2 CB 0.457 -0.823 -1.217
ALA 3 1HB 1.559 -0.896 -1.274
ALA 4 2HB 0.072 -1.860 -1.181
ALA 5 3HB 0.111 -0.382 -2.172
ARG 6 CA 0.000 0.000 0.000
ARG 7 CB 0.571 -1.447 0.034
ARG 8 CG 2.113 -1.554 0.199
ARG 9 CD 2.597 -3.014 0.186
ARG 10 NE 4.078 -3.055 0.368
ARG 11 CZ 4.818 -4.159 0.399
ARG 12 NH1 4.329 -5.359 0.264
ARG 13 NH2 6.096 -4.037 0.572
ARG 14 1HB 0.088 -2.009 0.858
ARG 15 2HB 0.249 -1.978 -0.886
ARG 16 1HG 2.623 -0.993 -0.612
ARG 17 2HG 2.415 -1.054 1.142
ARG 18 1HD 2.080 -3.581 0.990
ARG 19 2HD 2.301 -3.490 -0.775
ARG 20 HE 4.617 -2.191 0.487
ARG 21 1HH1 4.965 -6.157 0.299
ARG 22 2HH1 3.316 -5.385 0.129
ARG 23 1HH2 6.460 -3.088 0.675
ARG 24 2HH2 6.648 -4.897 0.594
ASN 25 CA 0.000 0.000 0.000
ASN 26 CB 0.566 -0.757 -1.237
ASN 27 CG 0.182 -2.234 -1.381
ASN 28 OD1 -0.701 -2.607 -2.139
ASN 29 ND2 0.806 -3.120 -0.652
ASN 30 1HB 0.233 -0.250 -2.164
ASN 31 2HB 1.668 -0.681 -1.265
ASN 32 1HD2 0.374 -4.046 -0.711
ASN 33 2HD2 1.473 -2.768 0.034
ASP 34 CA 0.000 0.000 0.000
ASP 35 CB 0.545 -0.744 -1.234
ASP 36 CG 2.087 -0.916 -1.305
ASP 37 OD1 2.711 -0.709 -0.261
ASP 38 OD2 2.508 -1.255 -2.412
ASP 39 1HB 0.151 -1.777 -1.244
ASP 40 2HB 0.194 -0.288 -2.174
CYS 41 CA 0.000 0.000 0.000
CYS 42 CB 0.553 -0.778 -1.210
CYS 43 SG 2.358 -0.827 -1.168
CYS 44 1HB 0.185 -1.820 -1.218
CYS 45 2HB 0.220 -0.328 -2.168
CYS 46 HG 2.565 0.065 -2.134
GLU 47 CA 0.000 0.000 0.000
GLU 48 CB 0.595 -0.719 -1.242
GLU 49 CG 0.204 -0.161 -2.651
```

```
GLU 50 CD 0.694 -0.911 -3.892
GLU 51 OE1 0.412 -0.567 -5.032
GLU 52 OE2 1.471 -1.993 -3.607
GLU 53 1HB 1.699 -0.716 -1.160
GLU 54 2HB 0.324 -1.792 -1.194
GLU 55 1HG -0.897 -0.095 -2.732
GLU 56 2HG 0.546 0.886 -2.754
GLU 57 HE1 0.606 0.438 -5.146
GLN 58 CA 0.000 0.000 0.000
GLN 59 CB 0.559 -0.738 -1.236
GLN 60 CG 2.117 -0.757 -1.384
GLN 61 CD 2.742 -1.419 -2.618
GLN 62 OE1 3.951 -1.419 -2.789
GLN 63 NE2 1.975 -1.974 -3.522
GLN 64 1HB 0.195 -1.785 -1.223
GLN 65 2HB 0.116 -0.287 -2.148
GLN 66 1HG 2.513 0.275 -1.395
GLN 67 2HG 2.578 -1.218 -0.491
GLN 68 1HE2 0.973 -1.880 -3.354
GLN 69 2HE2 2.467 -2.334 -4.342
GLY 70 CA 0.000 0.000 0.000
GLY 71 2HA 0.443 -0.488 -0.888
HIS 72 CA 0.000 0.000 0.000
HIS 73 CB 0.489 -0.703 -1.277
HIS 74 CG 1.394 -1.863 -0.978
HIS 75 ND1 1.777 -2.291 0.291
HIS 76 CD2 1.959 -2.649 -1.972
HIS 77 CE1 2.560 -3.334 -0.046
HIS 78 NE2 2.724 -3.612 -1.368
HIS 79 1HB -0.352 -1.084 -1.893
HIS 80 2HB 1.040 -0.002 -1.935
HIS 81 HD2 1.804 -2.501 -3.037
HIS 82 HE1 3.021 -3.894 0.758
HIS 83 HE2 3.279 -4.359 -1.810
ILE 84 CA 0.000 0.000 0.000
ILE 85 CB 0.677 -0.774 -1.203
ILE 86 CG1 0.285 -2.283 -1.310
ILE 87 CG2 2.234 -0.679 -1.185
ILE 88 CD1 0.658 -2.984 -2.633
ILE 89 HB 0.322 -0.285 -2.135
ILE 90 1HG1 0.694 -2.848 -0.450
ILE 91 2HG1 -0.813 -2.387 -1.212
ILE 92 1HG2 2.696 -1.173 -2.058
ILE 93 2HG2 2.607 0.360 -1.229
ILE 94 3HG2 2.669 -1.146 -0.281
ILE 95 1HD1 0.281 -4.023 -2.655
ILE 96 2HD1 0.226 -2.463 -3.509
ILE 97 3HD1 1.750 -3.041 -2.788
LEU 98 CA 0.000 0.000 0.000
LEU 99 CB 0.622 -0.783 -1.192
LEU 100 CG 0.335 -2.307 -1.279
LEU 101 CD1 0.822 -2.861 -2.628
LEU 102 CD2 1.008 -3.102 -0.146
```

```
LEU 103 1HB 0.297 -0.295 -2.133
LEU 104 2HB 1.721 -0.642 -1.183
LEU 105 HG -0.760 -2.468 -1.222
LEU 106 1HD1 0.600 -3.940 -2.733
LEU 107 2HD1 0.332 -2.354 -3.479
LEU 108 3HD1 1.914 -2.739 -2.760
LEU 109 1HD2 0.812 -4.187 -0.231
LEU 110 2HD2 2.106 -2.968 -0.138
LEU 111 3HD2 0.637 -2.795 0.849
LYS 112 CA 0.000 0.000 0.000
LYS 113 CB 0.615 -0.745 -1.222
LYS 114 CG 2.171 -0.719 -1.259
LYS 115 CD 2.809 -1.358 -2.500
LYS 116 CE 4.339 -1.174 -2.447
LYS 117 NZ 4.950 -1.800 -3.654
LYS 118 1HB 0.201 -0.307 -2.155
LYS 119 2HB 0.259 -1.794 -1.225
LYS 120 1HG 2.525 0.330 -1.234
LYS 121 2HG 2.564 -1.180 -0.329
LYS 122 1HD 2.519 -2.428 -2.551
LYS 123 2HD 2.382 -0.883 -3.409
LYS 124 1HE 4.598 -0.089 -2.382
LYS 125 2HE 4.749 -1.614 -1.507
LYS 126 1HZ 5.976 -1.735 -3.681
LYS 127 2HZ 4.726 -2.801 -3.733
LYS 128 3HZ 4.616 -1.375 -4.530
MET 129 CA 0.000 0.000 0.000
MET 130 CB 0.627 -0.721 -1.225
MET 131 CG 0.454 -2.254 -1.266
MET 132 SD 1.477 -2.949 -2.575
MET 133 CE 0.769 -4.601 -2.617
MET 134 1HB 0.220 -0.292 -2.163
MET 135 2HB 1.712 -0.504 -1.262
MET 136 1HG 0.729 -2.721 -0.302
MET 137 2HG -0.606 -2.512 -1.453
MET 138 1HE 0.818 -5.079 -1.622
MET 139 2HE 1.314 -5.238 -3.335
MET 140 3HE -0.290 -4.558 -2.930
PHE 141 CA 0.000 0.000 0.000
PHE 142 CB 0.614 -0.681 -1.261
PHE 143 CG 0.329 -2.180 -1.436
PHE 144 CD1 -0.823 -2.591 -2.117
PHE 145 CD2 1.172 -3.140 -0.870
PHE 146 CE1 -1.137 -3.942 -2.214
PHE 147 CE2 0.863 -4.493 -0.977
PHE 148 CZ -0.293 -4.894 -1.646
PHE 149 1HB 0.279 -0.144 -2.173
PHE 150 2HB 1.711 -0.531 -1.262
PHE 151 HD1 -1.500 -1.856 -2.533
PHE 152 HD2 2.058 -2.835 -0.331
PHE 153 HE1 -2.044 -4.247 -2.715
PHE 154 HE2 1.514 -5.231 -0.531
PHE 155 HZ -0.541 -5.942 -1.718
```

```
PRO 156 CA 0.000 0.000 0.000
PRO 157 CB 0.228 -1.399 -0.554
PRO 158 CG -1.024 -1.569 -1.431
PRO 159 CD -2.145 -0.959 -0.583
PRO 160 1HB 1.114 -1.391 -1.211
PRO 161 2HB 0.196 -2.125 0.268
PRO 162 1HG -1.061 -0.892 -2.300
PRO 163 2HG -1.136 -2.647 -1.607
PRO 164 1HD -2.936 -0.560 -1.240
PRO 165 2HD -2.506 -1.710 0.137
SER 166 CA 0.000 0.000 0.000
SER 167 CB 0.571 -0.809 -1.192
SER 168 OG 0.322 -0.187 -2.457
SER 169 1HB 1.666 -0.927 -1.072
SER 170 2HB 0.173 -1.842 -1.203
SER 171 HG -0.631 -0.158 -2.580
THR 172 CA 0.000 0.000 0.000
THR 173 CB 0.577 -0.785 -1.226
THR 174 OG1 0.204 -2.155 -1.162
THR 175 CG2 2.115 -0.809 -1.361
THR 176 HB 0.152 -0.345 -2.154
THR 177 HG1 0.527 -2.559 -1.973
THR 178 1HG2 2.447 -1.400 -2.235
THR 179 2HG2 2.544 0.201 -1.502
THR 180 3HG2 2.599 -1.246 -0.467
TRP 181 CA 0.000 0.000 0.000
TRP 182 CB 0.564 -0.745 -1.242
TRP 183 CG 0.356 -2.263 -1.229
TRP 184 CD1 -0.691 -2.954 -1.874
TRP 185 CD2 1.067 -3.218 -0.534
TRP 186 NE1 -0.646 -4.335 -1.601
TRP 187 CE2 0.452 -4.473 -0.766
TRP 188 CE3 2.198 -3.114 0.317
TRP 189 CZ2 0.971 -5.637 -0.158
TRP 190 CZ3 2.697 -4.278 0.902
TRP 191 CH2 2.094 -5.522 0.667
TRP 192 1HB 0.130 -0.322 -2.170
TRP 193 2HB 1.651 -0.558 -1.339
TRP 194 HD1 -1.458 -2.471 -2.465
TRP 195 HE1 -1.301 -5.064 -1.904
TRP 196 HE3 2.664 -2.158 0.510
TRP 197 HZ2 0.505 -6.595 -0.330
TRP 198 HZ3 3.562 -4.218 1.546
TRP 199 HH2 2.505 -6.407 1.130
TYR 200 CA 0.000 0.000 0.000
TYR 201 CB 0.583 -0.706 -1.259
TYR 202 CG 0.282 -2.201 -1.412
TYR 203 CD1 0.993 -3.151 -0.673
TYR 204 CD2 -0.720 -2.621 -2.292
TYR 205 CE1 0.699 -4.506 -0.810
TYR 206 CE2 -1.012 -3.975 -2.428
TYR 207 CZ -0.302 -4.916 -1.686
TYR 208 OH -0.589 -6.246 -1.815
```

```
TYR 209 1HB 0.250 -0.166 -2.170
TYR 210 2HB 1.684 -0.585 -1.277
TYR 211 HD1 1.768 -2.841 0.014
TYR 212 HD2 -1.284 -1.895 -2.863
TYR 213 HE1 1.242 -5.240 -0.234
TYR 214 HE2 -1.793 -4.285 -3.106
TYR 215 HH -1.298 -6.342 -2.453
VAL 216 CA 0.000 0.000 0.000
VAL 217 CB 0.619 -0.746 -1.248
VAL 218 CG1 0.430 -2.282 -1.217
VAL 219 CG2 2.136 -0.525 -1.501
VAL 220 HB 0.091 -0.369 -2.150
VAL 221 1HG1 0.982 -2.756 -0.384
VAL 222 2HG1 -0.633 -2.564 -1.107
VAL 223 3HG1 0.776 -2.757 -2.155
VAL 224 1HG2 2.760 -0.850 -0.647
VAL 225 2HG2 2.496 -1.068 -2.396
VAL 226 3HG2 2.381 0.537 -1.696
```

Bibliography

[1] Christian B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[2] Andrey Karshikoff. *Van der Waals Interactions*, chapter 2, pages 25–50. Imperial College Press; Distributed by World Scientific, 2006.

[3] Andrey Karshikoff. *Hydrophobic Interactions*, chapter 4, pages 91–128. Imperial College Press; Distributed by World Scientific, 2006.

[4] Arthur S. Edison. Linus pauling and the planar peptide bond. *Nat Struct Mol Biol*, 8(3):201–202, 03 2001.

[5] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666, 03 1958.

[6] Linus Pauling, Robert B Corey, and H R Branson. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, 37(4):205–211, 04 1951.

[7] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1):95 – 99, 1963.

[8] C B Anfinsen, E Haber, M Sela, and F H White. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, 47(9):1309–1314, 09 1961.

[9] Frederick H. White, D. Harker With a note by J. Bello, and E. De Jarnette. Regeneration of native secondary and tertiary structures by air oxidation of reduced ribonuclease. *Journal of Biological Chemistry*, 236(5):1353–1360, 1961.

[10] R Zwanzig, A Szabo, and B Bagchi. Levinthal's paradox. *Proceedings of the National Academy of Sciences of the United States of America*, 89(1):20–22, 01 1992.

[11] Andrej S'ali, Eugene Shakhnovich, and Martin Karplus. How does a protein fold? *Nature*, 369(6477):248–251, 05 1994.

[12] Ken A. Dill and Hue Sun Chan. From levinthal to pathways to funnels. *Nat Struct Mol Biol*, 4(1):10–19, 01 1997.

[13] Kunihiro Kuwajima, Katsutoshi Nitta, Michio Yoneyama, and Shintaro Sugai. Three-state denaturation of alpha-lactalbumin by guanidine hydrochloride. *Journal of Molecular Biology*, 106(2):359 – 373, 1976.

[14] Yash P. Myer. Conformation of cytochromes. III. effect of urea, temperature, extrinsic ligands, and ph variation on the conformation of horse heart ferricytochrome c. *Biochemistry*, 7(2):765–776, 1968. PMID: 5689422.

[15] Mikio Ohgushi and Akiyoshi Wada. molten-globule state: a compact form of globular proteins with mobile side-chains. *FEBS Letters*, 164(1):21–24, 1983.

[16] Emil Fischer. Einfluss der configuration auf die wirkung der enzyme. *Berichte der deutschen chemischen Gesellschaft*, 27(3):2985–2993, 1894.

[17] A E Mirsky and Linus Pauling. On the structure of native, denatured, and coagulated proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 22(7):439–447, 07 1936.

[18] Michael Sela, Frederick H. White, and Christian B. Anfinsen. Reductive cleavage of disulfide bridges in ribonuclease. *Science*, 125(3250):691–692, 1957.

[19] Raphael Zahler. Enzyme structure and mechanism. *The Yale Journal of Biology and Medicine*, 52(3):335–335, May-Jun 1979.

[20] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[21] Fred Karush. Heterogeneity of the binding sites of bovine serum albumin1. *Journal of the American Chemical Society*, 72(6):2705–2713, 1950.

[22] D E Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 44(2):98–104, 02 1958.

[23] W S Bennett and T A Steitz. Glucose-induced conformational change in yeast hexokinase. *Proceedings of the National Academy of Sciences of the United States of America*, 75(10):4848–4852, 10 1978.

[24] Colyn Crane-Robinson, Shirley E. Danby, E. Morton Bradbury, Annie Garel, Anne-Marie Kovacs, Madeleine Champagne, and Michel Daune. Structural studies of chicken erythrocyte histone h5. *European Journal of Biochemistry*, 67(2):379–388, 1976.

[25] Roland Riek, Simone Hornemann, Gerhard Wider, Martin Billeter, Rudi Glockshuber, and Kurt Wuthrich. Nmr structure of the mouse prion protein domain prp(121-231). *Nature*, 382(6587):180–182, 07 1996.

[26] C. Mark Fletcher and Gerhard Wagner. The interaction of eif4e with 4e-bp1 is an induced fit to a completely disordered protein. *Protein Science*, 7(7):1639–1642, 1998.

[27] Predrag Radivojac, Lilia M. Iakoucheva, Christopher J. Oldfield, Zoran Obradovic, Vladimir N. Uversky, and A. Keith Dunker. Intrinsic disorder and functional proteomics. *Biophysical Journal*, 92(5):1439 –1456, 2007.

[28] Megan Sickmeier, Justin A. Hamilton, Tanguy LeGall, Vladimir Vacic, Marc S. Cortese, Agnes Tantos, Beata Szabo, Peter Tompa, Jake Chen, Vladimir N. Uversky, Zoran Obradovic, and A. Keith Dunker. Disprot: the database of disordered proteins. *Nucleic Acids Research*, 35(suppl 1):D786–D793, 2007.

[29] Vladimir Vacic, Vladimir N. Uversky, A. Keith Dunker, and Stefano Lonardi. Composition profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics*, 8, 2007.

[30] S. Lise and D. T. Jones. Sequence patterns associated with disordered regions in proteins. *Proteins: Structure, Function, and Bioinformatics*, 58(1):144–150, 2005.

[31] John C. Wootton. Non-globular domains in protein sequences: Automated segmentation using complexity measures. *Computers & Chemistry*, 18(3):269 – 285, 1994.

[32] A.Keith Dunker, J.David Lawson, Celeste J Brown, Ryan M Williams, Pedro Romero, Jeong S Oh, Christopher J Oldfield, Andrew M Campen, Catherine M Ratliff, Kerry W Hipps, Juan Ausio, Mark S Nissen, Raymond Reeves, ChulHee Kang, Charles R Kissinger, Robert W Bailey, Michael D Griswold, Wah Chiu, Ethan C Garner, and

Zoran Obradovic. Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling*, 19(1):26 – 59, 2001.

[33] A. Keith Dunker, Celeste J. Brown, J. David Lawson, Lilia M. Iakoucheva, and Zoran Obradovi. Intrinsic disorder and protein function. *Biochemistry*, 41(21):6573–6582, 2002. PMID: 12022860.

[34] Lars Bertram and Rudolph E. Tanzi. Alzheimer's disease: one disorder, too many genes? *Human Molecular Genetics*, 13(suppl-1):R135, 2004.

[35] Mark Wells, Henning Tidow, Trevor J. Rutherford, Phineus Markwick, Malene Ringkjobing Jensen, Efstratios Mylonas, Dmitri I. Svergun, Martin Blackledge, and Alan R. Fersht. Structure of tumor suppressor p53 and its intrinsically disordered n-terminal transactivation domain. *Proceedings of the National Academy of Sciences*, 105(15):5762–5767, 2008.

[36] Peter E. Wright and H. Jane Dyson. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol*, 16(1):18–29, 01 2015.

[37] H.Jane Dyson and Peter E Wright. Coupling of folding and binding for unstructured proteins. *Current Opinion in Structural Biology*, 12(1):54 – 60, 2002.

[38] Peter Tompa and Monika Fuxreiter. Fuzzy complexes: polymorphism and structural disorder in proteinprotein interactions. *Trends in Biochemical Sciences*, 33(1):2 – 8, 2008.

[39] Alexander B Sigalov, Anastasia V Zhuravleva, and Vladislav Yu Orekhov. Binding of intrinsically disordered proteins is not necessarily accompanied by a structural transition to a folded form. *Biochimie*, 89(3):419–421, 03 2007.

[40] Alexander B Sigalov, Walter M Kim, Maria Saline, and Lawrence J Stern. Intrinsically disordered cytoplasmic domain of t cell receptor zeta chain binds to the nef protein of simian immunodeficiency virus without a disorder-to-order transition(). *Biochemistry*, 47(49):12942–12944, 12 2008.

[41] Rajendrani Mukhopadhyay and Jan H. Hoh. Afm force measurements on microtubule-associated proteins: the projection domain exerts a long-range repulsive force. *FEBS Letters*, 505(3):374–378, 2001.

[42] Huan-Xiang Zhou. The affinity-enhancing roles of flexible linkers in two-domain dna-binding proteins. *Biochemistry*, 40(50):15069–15073, 2001. PMID: 11735389.

[43] Wolfgang A. Linke, Michael Kulke, Hongbin Li, Setsuko Fujita-Becker, Ciprian Neagoe, Dietmar J. Manstein, Mathias Gautel, and Julio M. Fernandez. Pevk domain of titin: An entropic spring with actin-binding properties. *Journal of Structural Biology*, 137(1):194 – 205, 2002.

[44] Jeffrey G Forbes, Albert J Jin, Kan Ma, Gustavo Gutierrez-Cruz, Wanxia L Tsai, and Kuan Wang. Titin pevk segment: charge-driven elasticity of the open and flexible polyampholyte. *Journal of muscle research and cell motility*, 26(0):291–301, 2005.

[45] Diana Ekman, Sara Light, Åsa K Björklund, and Arne Elofsson. What properties characterize the hub proteins of the protein-protein interaction network of saccharomyces cerevisiae? *Genome Biology*, 7(6):R45–R45, 2006.

[46] Vladimir N. Uversky. The multifaceted roles of intrinsic disorder in protein complexes. *FEBS Letters*, 589(19PartA):2498–2506, 2015.

[47] Ursula Jakob, Richard Kriwacki, and Vladimir N. Uversky. Conditionally and transiently disordered proteins: Awakening cryptic disorder to regulate protein function. *Chemical Reviews*, 114(13):6779–6805, 2014. PMID: 24502763.

[48] Sergei E. Permyakov, Ian S. Millett, Sebastian Doniach, Eugene A. Permyakov, and Vladimir N. Uversky. Natively unfolded c-terminal domain of caldesmon remains substantially unstructured after the effective binding to calmodulin. *Proteins: Structure, Function, and Bioinformatics*, 53(4):855–n/a, 2003.

[49] Alexander B. Sigalov, Anastasia V. Zhuravleva, and Vladislav Yu. Orekhov. Binding of intrinsically disordered proteins is not necessarily accompanied by a structural transition to a folded form. *Biochimie*, 89(3):419 – 421, 2007.

[50] Sascha Martens and Harvey T. McMahon. Mechanisms of membrane fusion: disparate players and common principles. *Nat Rev Mol Cell Biol*, 9(7):543–556, 07 2008.

[51] Thomas C. Südhof and James E. Rothman. Membrane fusion: Grappling with snare and sm proteins. *Science*, 323(5913):474–477, 2009.

[52] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[53] John F. Marko and Eric D. Siggia. Stretching DNA. *Macromolecules*, 28(26):8759–8770, 1995.

[54] Paul J. Flory. The configuration of real polymer chains. *The Journal of Chemical Physics*, 17(3):303–310, 1949.

[55] Arthur S. Tatham and Peter R. Shewry. Elastomeric proteins: biological roles, structures and mechanisms. *Trends in Biochemical Sciences*, 25(11):567 – 571, 2000.

[56] Mark B van Eldijk, Christopher L McGann, Kristi L Kiick, and Jan CM van Hest. Elastomeric polypeptides. *Topics in current chemistry*, 310:71–116, 2012.

[57] M D Wang, H Yin, R Landick, J Gelles, and S M Block. Stretching dna with optical tweezers. *Biophysical Journal*, 72(3):1335–1346, 03 1997.

[58] Guillaume Stirnemann, David Giganti, Julio M. Fernandez, and B. J. Berne. Elasticity, structure, and relaxation of extended proteins under force. *Proceedings of the National Academy of Sciences*, 110(10):3847–3852, 2013.

[59] Piotr E. Marszalek, Hui Lu, Hongbin Li, Mariano Carrion-Vazquez, Andres F. Oberhauser, Klaus Schulten, and Julio M. Fernandez. Mechanical unfolding intermediates in titin modules. *Nature*, 402(6757):100–103, 11 1999.

[60] C Bouchiat, M D Wang, J Allemand, T Strick, S M Block, and V Croquette. Estimating the persistence length of a worm-like chain molecule from force-extension measurements. *Biophysical Journal*, 76(1 Pt 1):409–413, 01 1999.

[61] C. A. J. Hoeve and P. J. Flory. The elastic properties of elastin. *Biopolymers*, 13(4):677–686, 1974.

[62] John M. Gosline. Hydrophobic interaction and a model for the elasticity of elastin. *Biopolymers*, 17(3):677–695, 1978.

[63] M. Hong, D. Isailovic, R.A. McMillan, and V.P. Conticello. Structure of an elastin-mimetic polypeptide by solid-state nmr chemical shift analysis. *Biopolymers*, 70(2):158–168, 2003.

[64] D. W. Urry, T. L. Trapane, and K. U. Prasad. Phase-structure transitions of the elastin polypentapeptidewater system within the framework of compositiontemperature studies. *Biopolymers*, 24(12):2345–2356, 1985.

[65] Faqir Muhammad Anjum, Moazzam Rafiq Khan, Ahmad Din, Muhammad Saeed, Imran Pasha, and Muhammad Umair Arshad. Wheat gluten: High molecular weight glutenin subunitsstructure, genetics, and relation to dough elasticity. *Journal of Food Science*, 72(3):R56–R63, 2007.

[66] M. C. Gianibelli, O. R. Larroque, F. MacRitchie, and C. W. Wrigley. Biochemical, genetic, and molecular characterization of wheat glutenin and its component subunits. *Cereal Chemistry Journal*, 78(6):635–646, 2017/02/04 2001.

[67] Peter R Shewry, Arthur S Tatham, and Paul Lazzeri. Biotechnology of wheat quality. *Journal of the Science of Food and Agriculture*, 73(4):397–406, 1997.

[68] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wirkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B*, 102(18):3586–3616, 1998.

[69] Alexander D. MacKerell, Michael Feig, and Charles L. Brooks. Improved treatment of the protein backbone in empirical force fields. *Journal of the American Chemical Society*, 126(3):698–699, 2004. PMID: 14733527.

[70] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983.

[71] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kal, and Klaus Schul-

ten. Scalable molecular dynamics with namd. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[72] Dmitrij Frishman and Patrick Argos. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics*, 23(4):566–579, 1995.

[73] Jonathan E. Kohn, Ian S. Millett, Jaby Jacob, Bojan Zagrovic, Thomas M. Dillon, Nikolina Cingel, Robin S. Dothager, Soenke Seifert, P. Thiyagarajan, Tobin R. Sosnick, M. Zahid Hasan, Vijay S. Pande, Ingo Ruczinski, Sebastian Doniach, and Kevin W. Plaxco. Random-coil behavior and the dimensions of chemically unfolded proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 101(34):12491–12496, 2004.

[74] Sri Rama Koti Ainavarapu, Jasna Brujić, Hector H Huang, Arun P Wiita, Hui Lu, Lewyn Li, Kirstin A Walther, Mariano Carrion-Vazquez, Hongbin Li, and Julio M Fernandez. Contour length and refolding rate of a small protein controlled by engineered disulfide bonds. *Biophysical Journal*, 92(1):225–233, 01 2007.

[75] Alexei Valiaev, Dong Woo Lim, Scott Schmidler, Robert L. Clark, Ashutosh Chilkoti, and Stefan Zauscher. Hydration and conformational mechanics of single, end-tethered elastin-like polypeptides. *Journal of the American Chemical Society*, 130(33):10939–10946, 2008. PMID: 18646848.

[76] Theo Odijk. Stiff chains and filaments under tension. *Macromolecules*, 28(20):7016–7018, 1995.

[77] Robin van der Lee, Marija Buljan, Benjamin Lang, Robert J. Weatheritt, Gary W. Daughdrill, A. Keith Dunker, Monika Fuxreiter, Julian Gough, Joerg Gsponer, David T. Jones, Philip M. Kim, Richard W. Kriwacki, Christopher J. Oldfield, Rohit V. Pappu, Peter Tompa, Vladimir N. Uversky, Peter E. Wright, and M. Madan

Babu. Classification of intrinsically disordered regions and proteins. *Chemical Reviews*, 114(13):6589–6631, 2014. PMID: 24773235.

[78] Nick J. Proudfoot, Andre Furger, and Michael J. Dye. Integrating mrna processing with transcription. *Cell*, 108(4):501–512, 2017/03/10.

[79] Patricia M. Okamoto, Jonathan S. Herskovits, and Richard B. Vallee. Role of the basic, proline-rich region of dynamin in src homology 3 domain binding and endocytosis. *Journal of Biological Chemistry*, 272(17):11629–11635, 1997.

[80] Juan S Bonifacino and Benjamin S Glick. The mechanisms of vesicle budding and fusion. *Cell*, 116(2):153 – 166, 2004.

[81] Gerrit J. K. Praefcke and Harvey T. McMahon. The dynamin superfamily: universal membrane tubulation and fission molecules? *Nat Rev Mol Cell Biol*, 5(2):133–147, 02 2004.

[82] Ammon E. Posey. *Intrinsic disorder and amphitropism in the mitochondrial fission mechanoenzyme drp1*. PhD thesis, Johns Hopkins University, 2014.

[83] Brian J. Bennion and Valerie Daggett. Counteraction of urea-induced protein denaturation by trimethylamine n-oxide: A chemical chaperone at atomic resolution. *Proceedings of the National Academy of Sciences of the United States of America*, 101(17):6433–6438, 2004.

[84] Ilia Baskakov and D. Wayne Bolen. Forcing thermodynamically unfolded proteins to fold. *Journal of Biological Chemistry*, 273(9):4831–4834, 1998.

[85] Char Y. Hu, Gillian C. Lynch, Hironori Kokubo, and B. Montgomery Pettitt. Trimethylamine n-oxide influence on the backbone of proteins: An oligoglycine model. *Proteins: Structure, Function, and Bioinformatics*, 78(3):695–704, 2010.

[86] Robert B. Best, Xiao Zhu, Jihyun Shim, Pedro E. M. Lopes, Jeetain Mittal, Michael Feig, and Alexander D. MacKerell. Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone ϕ, *psi* and side-chain 1 and 2 dihedral angles. *Journal of Chemical Theory and Computation*, 8(9):3257–3273, 2012. PMID: 23341755.

[87] Kristine M. Kast, Jürgen Brickmann, Stefan M. Kast, and R. Stephen Berry. Binary phases of aliphatic n-oxides and water: force field development and molecular dynamics simulation. *The Journal of Physical Chemistry A*, 107(27):5342–5351, 2003.

[88] Deepak R. Canchi, Pruthvi Jayasimha, Donald C. Rau, George I. Makhatadze, and Angel E. Garcia. Molecular mechanism for the preferential exclusion of tmao from protein surfaces. *The Journal of Physical Chemistry B*, 116(40):12095–12104, 2012. PMID: 22970901.

[89] G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187 – 199, 1977.

[90] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992.

[91] Fangqiang Zhu and Gerhard Hummer. Convergence and error estimation in free energy calculations using the weighted histogram analysis method. *Journal of Computational Chemistry*, 33(4):453–465, 2012.

[92] M. Rubinstein and R.H. Colby. *Polymer Physics*. OUP Oxford, 2003.

[93] Deborah K. Wilkins, Shaun B. Grimshaw, Vronique Receveur, Christopher M. Dobson, Jonathan A. Jones, and Lorna J. Smith. Hydrodynamic radii of native and denatured proteins measured by pulse field gradient nmr techniques. *Biochemistry*, 38(50):16424–16431, 1999. PMID: 10600103.

[94] Samuel S. Cho, Govardhan Reddy, John E. Straub, and D. Thirumalai. Entropic stabilization of proteins by tmao. *The Journal of Physical Chemistry B*, 115(45):13401–13407, 2011. PMID: 21985427.

[95] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell. Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. *Journal of Computational Chemistry*, 31(4):671–690, 2010.

[96] Shashi Prajapati, Vinod Bhakuni, Kodali Ravindra Babu, and Swatantra Kumar Jain. Alkaline unfolding and salt-induced folding of bovine liver catalase at high ph. *European Journal of Biochemistry*, 255(1):178–184, 1998.

[97] Marlene Andersson, Jan Johansson, and Anna Rising. Silk spinning in silkworms and spiders. *International Journal of Molecular Sciences*, 17(8):1290, 08 2016.

[98] Chengjie Fu, Zhengzhong Shao, and Vollrath Fritz. Animal silks: their structures, properties and artificial production. *Chem. Commun.*, pages 6515–6529, 2009.

[99] Fiorenzo G. Omenetto and David L. Kaplan. New opportunities for an ancient material. *Science*, 329(5991):528–531, 2010.

[100] Joydip Kundu, Moumita Dewan, Sarani Ghoshal, and S. C. Kundu. Mulberry non-engineered silk gland protein vis-à-vis silk cocoon protein engineered by silkworms as

biomaterial matrices. *Journal of Materials Science: Materials in Medicine*, 19(7):2679–2689, 2008.

[101] Yang Cao and Bochu Wang. Biodegradation of silk biomaterials. *International Journal of Molecular Sciences*, 10(4):1514–1524, 04 2009.

[102] Yong-Xing He, Nan-Nan Zhang, Wei-Fang Li, Ning Jia, Bao-Yu Chen, Kang Zhou, Jiahai Zhang, Yuxing Chen, and Cong-Zhao Zhou. N-terminal domain of bombyx mori fibroin mediates the assembly of silk in response to ph decrease. *Journal of Molecular Biology*, 418(34):197 – 207, 2012.

[103] Andrew A. Walker, Chris Holland, and Tara D. Sutherland. More than one way to spin a crystallite: multiple trajectories through liquid crystallinity to solid silk. *Proceedings of the Royal Society of London B: Biological Sciences*, 282(1809), 2015.

[104] Kazuei Mita, Sachiko Ichimura, and Tharappel C. James. Highly repetitive structure and its organization of the silk fibroin gene. *Journal of Molecular Evolution*, 38(6):583–592, 1994.

[105] Bernard Lotz and Franois Colonna Cesari. The chemical structure and the crystalline structures of bombyx mori silk fibroin. *Biochimie*, 61(2):205 – 214, 1979.

[106] Richard E. Marsh, Robert B. Corey, and Linus Pauling. An investigation of the structure of silk fibroin. *Biochimica et Biophysica Acta*, 16:1 – 34, 1955.

[107] Tetsuo Asakura, Kosuke Ohgo, Teppei Ishida, Paola Taddei, Patrizia Monti, and Raghuvansh Kishore. Possible implications of serine and tyrosine residues and intermolecular interactions on the appearance of silk i structure of bombyx mori silk fibroin-derived synthetic peptides: high-resolution 13c cross-polarization/magic-angle spinning nmr study. *Biomacromolecules*, 6(1):468–474, 2005. PMID: 15638554.

[108] Hazime Saito, Mika Ishida, Motoko Yokoi, and Tetsuo Asakura. Dynamic features of side chains in tyrosine and serine residues of some polypeptides and fibroins in the solid as studied by high-resolution solid-state carbon-13 nmr spectroscopy. *Macromolecules*, 23(1):83–88, 1990.

[109] Shu-ichi Nakano, Hirohito Oka, Yuuki Uotani, Kazuya Uenishi, Masayuki Fujii, and Naoki Sugimoto. Stacking interaction in the middle and at the end of a dna helix studied with non-natural nucleotides. *Mol. BioSyst.*, 6:2023–2029, 2010.

[110] Jonathan R. Lai, Bayard R. Huck, Bernard Weisblum, and Samuel H. Gellman. Design of non-cysteine-containing antimicrobial -hairpins: structureactivity relationship studies with linear protegrin-1 analogues. *Biochemistry*, 41(42):12835–12842, 2002. PMID: 12379126.

[111] Esteban Lanzarotti, Rolf R. Biekofsky, Daro A. Estrin, Marcelo A. Marti, and Adrin G. Turjanski. Aromaticaromatic interactions in proteins: Beyond the dimer. *Journal of Chemical Information and Modeling*, 51(7):1623–1633, 2011. PMID: 21662246.

[112] Benjamin P. Partlow, Mehran Bagheri, James L. Harden, and David L. Kaplan. Tyrosine templating in the self-assembly and crystallization of silk fibroin. *Biomacromolecules*, 17(11):3570–3579, 2016. PMID: 27736062.

[113] Benjamin P. Partlow, Craig W. Hanna, Jelena Rnjak-Kovacina, Jodie E. Moreau, Matthew B. Applegate, Kelly A. Burke, Benedetto Marelli, Alexander N. Mitropoulos, Fiorenzo G. Omenetto, and David L. Kaplan. Highly tunable elastomeric silk biomaterials. *Advanced Functional Materials*, 24(29):4615–4624, 2014.

[114] David J. Raven, Christopher Earland, and Marjorie Little. Occurrence of dityrosine in tussah silk fibroin and keratin. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 251(1):96 – 99, 1971.

[115] Ching-Shuen Wang, Nicholas N. Ashton, Robert B. Weiss, and Russell J. Stewart. Peroxinectin catalyzed dityrosine crosslinking in the adhesive underwater silk of a case-maker caddisfly larvae, hysperophylax occidentalis. *Insect Biochemistry and Molecular Biology*, 54:69 – 79, 2014.

[116] Jos Roberto Aparecido dos Santos-Pinto, Gnther Lamprecht, Wei-Qiang Chen, Seok Heo, John George Hardy, Helga Priewalder, Thomas Rainer Scheibel, Mario Sergio Palma, and Gert Lubec. Structure and post-translational modifications of the web silk protein spidroin-1 from nephila spiders. *Journal of Proteomics*, 105:174 – 185, 2014. Special Issue: Proteomics of non-model organisms.

[117] Sinan Keten and Markus J Buehler. Nanostructure and molecular mechanics of spider dragline silk protein assemblies. *Journal of the Royal Society Interface*, 7(53):1709–1721, 12 2010.

[118] Andrew A. Walker, Chris Holland, and Tara D. Sutherland. More than one way to spin a crystallite: multiple trajectories through liquid crystallinity to solid silk. *Proceedings of the Royal Society of London B: Biological Sciences*, 282(1809), 2015.

[119] Hyoung-Joon Jin and David L. Kaplan. Mechanism of silk processing in insects and spiders. *Nature*, 424(6952):1057–1061, 08 2003.

[120] Irene Georgakoudi, Irene Tsai, Cherry Greiner, Cheryl Wong, Jordy DeFelice, and David Kaplan. Intrinsic fluorescence changes associated with the conformational state of silk fibroin in biomaterial matrices. *Opt. Express*, 15(3):1043–1053, Feb 2007.

[121] Anne Martel, Manfred Burghammer, Richard J. Davies, Emanuela Di Cola, Charlotte Vendrely, and Christian Riekel. Silk fiber assembly studied by synchrotron radiation saxs/waxs and raman spectroscopy. *Journal of the American Chemical Society*, 130(50):17070–17074, 2008. PMID: 19053481.

[122] Georgia B. McGaughey, Marc Gagn, and Anthony K. Rapp. -stacking interactions: Alive and well in proteins. *Journal of Biological Chemistry*, 273(25):15458–15463, 1998.

[123] Jane S. Merkel and Lynne Regan. Aromatic rescue of glycine in beta-sheets. *Folding and Design*, 3(6):449 – 456, 1998.

[124] Tetsuo Asakura, Kohei Suita, Tsunenori Kameda, Sergii Afonin, and Anne S. Ulrich. Structural role of tyrosine in bombyx mori silk fibroin, studied by solid-state nmr and molecular mechanics on a model peptide prepared as silk i and ii. *Magnetic Resonance in Chemistry*, 42(2):258–266, 2004.

[125] Tetsuo Asakura, Keiko Okushita, and Mike P. Williamson. Analysis of the structure of bombyx mori silk fibroin by nmr. *Macromolecules*, 48(8):2345–2357, 2015.

[126] Paola Taddei, Tetsuo Asakura, Juming Yao, and Patrizia Monti. Raman study of poly(alanine-glycine)-based peptides containing tyrosine, valine, and serine as model for the semicrystalline domains of bombyx mori silk fibroin. *Biopolymers*, 75(4):314–324, 2004.

[127] S. Rammensee, U. Slotta, T. Scheibel, and A. R. Bausch. Assembly mechanism of recombinant spider silk proteins. *Proceedings of the National Academy of Sciences*, 105(18):6590–6595, 2008.

[128] Marie-Eve Rousseau, Thierry Lefvre, Lilyane Beaulieu, Tetsuo Asakura, and Michel Pzolet. Study of protein conformation and orientation in silkworm and spider silk fibers using raman microspectroscopy. *Biomacromolecules*, 5(6):2247–2257, 2004. PMID: 15530039.

[129] Cong-Zhao Zhou, Fabrice Confalonieri, Michel Jacquet, Roland Perasso, Zhen-Gang Li, and Joel Janin. Silk fibroin: Structural implications of a remarkable amino acid sequence. *Proteins: Structure, Function, and Bioinformatics*, 44(2):119–122, 2001.