THE UNIVERSITY OF CHICAGO

MOLECULAR SIMULATION OF BIOLOGICAL MACROMOLECULES: AGGREGATION, ION CAPTURE, AND SLIDING DYNAMICS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE PRITZKER SCHOOL OF MOLECULAR ENGINEERING
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
ASHLEY ZHANG GUO

CHICAGO, ILLINOIS
JUNE 2020

TABLE OF CONTENTS

LIST OF FIGURES v

LIST OF TABLES x

ACKNOWLEDGMENTS xi

ABSTRACT xiii

1 INTRODUCTION 1

2 EARLY-STAGE HUMAN ISLET AMYLOID POLYPEPTIDE AGGREGATION: MECHANISMS BEHIND DIMER FORMATION 5
2.1 Introduction 5
2.2 Results and Discussion 8
2.2.1 Free Energy Landscape for the hIAPP Dimer 8
2.2.2 hIAPP Dimerization Pathway 11
2.3 Conclusions 17
2.4 Methods 18
2.4.1 Human Amylin Dimer 18
2.4.2 Bias-Exchange Metadynamics 18
2.4.3 Finite Temperature String Method 21

3 MOLECULAR INSIGHTS INTO THE ROLE OF WATER IN EARLY-STAGE HUMAN AMYLIN AGGREGATION 24
3.1 Introduction 24
3.2 Results and Discussion 27
3.2.1 Trimer Assembly from the Disordered State (“3-chain Assembly”) . . 27
3.2.2 Trimer Assembly from the Dimer State (“2+1 Assembly”) 29
3.2.3 “3-chain Assembly” versus “2+1 Assembly” 35
3.3 Conclusions 40
3.4 Methods 41
3.4.1 Human Amylin Trimer 41
3.4.2 Finite Temperature String Method 42

4 DYNAMICS OF PEPTIDE AMPHIPHILES FOR PHOSPHATE CAPTURE AND RELEASE 45
4.1 Introduction 45
4.2 Results and Discussion 48
4.2.1 Phosphate Binding to a Single C16GGGhex Chain 48
4.2.2 Phosphate Binding to a Flat Layer of C16GGGhex 52
4.3 Conclusions 55
4.4 Methods 57

5 EXTRACTING COLLECTIVE MOTIONS UNDERLYING NUCLEOSOME DYNAMICS VIA NONLINEAR MANIFOLD LEARNING 59
5.1 Introduction 59
5.2 Methods 62
5.2.1 MD Simulations of the Nucleosome 62
5.2.2 Diffusion maps 62
5.2.3 Collective Variables Describing DNA Translocation and Rotation . . 64
5.2.4 Collective Variable Describing DNA Breathing 65
5.2.5 Collective Variable Describing DNA Looping 67
5.3 Results and Discussion 68
5.3.1 DNA Translocation 69
5.3.2 DNA Rotation 70
5.3.3 DNA Breathing 71
5.3.4 DNA Looping 73
5.4 Conclusions 76

6 CONCLUSIONS 79

REFERENCES 82

LIST OF FIGURES

2.1 (a) Free energy landscape of the hIAPP dimer as a function of $Q_{\text{res 8-16}}$ and $Q_{\text{res 27-35}}$, obtained via BE-MetaD and the AMBERff99SB*-ILDN force field. Free energy values are relative to the global minimum at $(Q_{\text{res 8-16}}, Q_{\text{res 27-35}}) = (0.43, 0.38)$. Contour lines are plotted for free energies from 2.5 kJ/mol (black) to 15 kJ/mol (green), with stride 2.5 kJ/mol. All free energies greater than or equal to 15 kJ/mol are plotted in yellow. Two local minima corresponding to dimer structures are located, at $(0.88, 0.88)$ and $(0.88, 0.68)$. (b) Average amount of parallel β-sheet spanning residues 20-29 in the hIAPP dimer in solution, as measured by $\beta^{\text{parallel}}_{\text{RMSD 20-29}}$, as function of $Q_{\text{res 8-16}}$ and $Q_{\text{res 27-35}}$. The value of $\beta^{\text{parallel}}_{\text{RMSD 20-29}}$ ranges from 0 (no parallel β-sheet) to 8 (completely parallel β-sheet). Statistics were collected from the full production simulation time of all six BE-MetaD replicas. Free energy contours from (a) are superimposed on the plot. 9

2.2 Dimerization pathway obtained via finite temperature string method. Five representative snapshots illustrate conformational changes that take place during dimer formation. Water and counterions are not shown for clarity. The 35-node pathway is composed of the results of a 32-node and smaller 4-node string method, as described in Methods. Grey points show initial configurations for the 32-node and 4-node string method calculations. 12

2.3 Free energy profile along the dimerizaton pathway obtained via finite temperature string method. The reaction coordinate proceeds from 0.0 (disordered state) to 1.0 (fully formed dimer). Free energy is calculated as described in Methods, sampling 35 Voronoi cells for 150 ns each. Representative snapshots illustrate key conformational changes, including the formation of a transient β-sheet structure after surpassing the initial free energy barrier of approximately 7 k_BT. Average secondary structure per residue across the entire dimerization pathway is presented in Figure 2.4. 13

2.4 Average β-sheet and α-helix secondary structure at each of the points where a free energy calculation was performed in Figure 2.3, plotted by residue (ranging from 1 to 37) versus reaction coordinate (from the disordered state at 0.0 to fully dimerized state at 1.0). Secondary structure was assigned using the DSSP algorithm [53] and averaged over all 150 ns of sampling for each bin. The Barrier I and II regions corresponding to the barriers highlighted in Figure 2.3 are marked in both plots. As Barrier I is crossed, a transient β-sheet is formed in residues 19-24 and in the L12A13 region. The crossing of Barrier II corresponds to simultaneous loss of β-sheet in the C-termini and increased β-sheet in the N-termini, with regions of high β-sheet character gradually moving out towards the termini as the full dimer is formed, accompanied with a drop in free energy. 14

2.5 Free energy decomposition into enthalpic (ΔU) and entropic ($-\Delta TS$) contributions for dimerization. ΔA is taken from the free energy calculation, performed as described in Methods. Average potential energy is calculated from each of the 35 Voronoi cells sampled during free energy calculation, from a total of 15003 snapshots per cell. Entropic contributions are then calculated as $-\Delta TS = \Delta A - \Delta U$. 16

2.6 Six RMSD-aligned snapshots of the dimer, taken from the trajectory corresponding to the free energy barrier marked by snapshot E in Figure 2. While the core of the dimer structure remains compact across the multiple snapshots, the termini of each hIAPP molecule explore more freely, contributing to the jump in entropy found in Figure 2.5. Structures were aligned using VMD [9, 49]. 16

3.1 3-chain trimerization pathway calculated from the finite temperature string method, with initial configurations input into the string method shown in grey. Four representative snapshots show mechanistic details during trimer formation. Water and counterions are removed for clarity. The disordered end of the string (Panel I) includes the previously studied dimer intermediate (formed by the yellow and red chains). A similar intermediate comprised of all three chains is discovered and shown in Panel II. 28

3.2 Free energy profile along the 3-chain trimerization pathway found via the finite temperature string method. The reaction coordinate extends from 0.0 (disordered state) to 1.0 (fully formed trimer). Free energy is calculated using the procedure described in the Methods, sampling each of the 32 cells for 150 ns each. Representative snapshots are shown to illustrate conformational changes during trimerization. A free energy barrier of approximately 12 k_BT is found, corresponding to the formation of a transient β-sheet structure (shown in Panel C). 30

3.3 Average α-helix and β-sheet secondary structure along the 3-chain trimerization pathway, plotted by residue (ranging from 1 to 37, averaged over all three hIAPP chains). Secondary structure was calculated using the DSSP algorithm[53] and averaged over 150 ns of sampling for each of the 32 bins along the reaction coordinate. β-sheet is found to transiently form in residues 12–13 and 20–29 during formation of the intermediate β-sheet structure, and as trimerization progresses, β-sheet extends to both termini, with more heavily β-sheet character observed in the C-termini. 31

3.4 2+1 assembly pathway calculated from the finite temperature string method. Grey points show initial configurations input into the string method. Four representative snapshots illustrate conformational changes observed during trimer formation. Water and counterions are not shown. The disordered end of the string (Panel I) shows a loosely formed dimer and a disordered third chain. Increased β-sheet is formed gradually until the full trimer is formed (Panel IV). . 32

3.5 Free energy profile along the 2+1 assembly pathway found via the finite temperature string method. The reaction coordinate proceeds from 0.0 (disordered state containing a loosely formed dimer and third disordered chain) to 1.0 (fully formed trimer). Free energy is calculated using the procedure described in the Methods, sampling each of the 32 cells for 150 ns each. Representative snapshots are shown to illustrate conformational changes during trimerization. A free energy barrier of approximately 40 k_BT is found. The intermediate transient β-sheet structure observed during 3-chain assembly is not found. 33

3.6 Average α-helix and β-sheet secondary structure along the 2+1 assembly pathway, plotted by residue (ranging from 1 to 37, averaged over all three hIAPP chains). Secondary structure was calculated using the DSSP algorithm[53] and averaged over 150 ns of sampling for each of the 32 bins along the reaction coordinate. As with 3+1 assembly, β-sheet extends to both termini, with more heavily β-sheet character observed in the C-termini. 34

3.7 Average protein-protein and protein-water hydrogen bonds within 3 Åover the course of dimerization for both trimerization processes, shown with standard error. Average number of H-bonds are calculated from each of the 32 cells sampled during free energy calculation, from 15003 snapshots per cell, using the GROMACS g_hbond tool. While protein-protein H-bonds trend upward during both assembly processes, the protein-water H-bonds show a steeper decrease in 3-chain assembly compared to 2+1 assembly. Protein-protein H-bonds are found to form while protein-water H-bonds are lost during 3-chain assembly; this is not observed for 2+1 assembly. 36

3.8 Decomposition of free energy into enthalpic (ΔU) and entropic ($-\Delta TS$) contributions for both trimerization mechanisms. ΔA is taken from the free energy calculation, performed as described in Methods and shown in Figures 3.2 and 3.5. Average potential energy is calculated from each of the 32 cells sampled during free energy calculation, from a total of 15003 snapshots per cell. Entropic contributions are then calculated as $-\Delta TS = \Delta A - \Delta U$. 37

3.9 Decomposition of potential energy into interchain, intrachain, and chain-water interactions for both trimerization mechanisms. Energies are calculated from each of the 32 cells sampled during free energy calculation, from a total of 15003 snapshots per cell. Note the contrasting trends between the two assembly processes for all three interactions plotted. 38

3.10 Comparison of free energy profiles for both 3-chain and 2+1 assembly, using an explicit water model (as was shown in Figures 3.2 and3.5) versus implicit water. When the explicit water interactions are replaced with a dielectric continuum as in the implicit model, the free energy profiles are noticeably flattened, suggesting that the moleuclar-level interactions with water are responsible for the free energy barriers originally observed. 39

4.1 Molecular structure of C16GGGhex. A C16 alkyl tail is connected by a 3 glycine linker to the SGAGKT hexapeptide, which has previously been demonstrated to bind to phosphate in a pH-responsive manner. 48

4.2 Top 8 most occupied clusters obtained for both high and low pH single-chain phosphate binding simulations. Populations of clusters decrease from left to right. Note the multiple locations at which phosphate is found to associate with the C16GGGhex chain; this motivates our choice in collective variables described in the text. 50

4.3 Free energy surface for single-chain phosphate binding at low pH conditions. Three distinct free energy minima are found for the low pH system, with moderate free energy barriers between them. 51

4.4 Free energy surface for single-chain phosphate binding at high pH conditions. Unlike the low pH free energy landscape in Figure 4.3, there are no clear minima associated with bound phosphate. Free energy steadily decreases as the phosphate moves further away from the C16GGGhex chain. 52

4.5 Schematic showing setup of flat layer C16GGGhex simulations. 25 C16GGGhex chains are distributed evenly across the x-y plane of the box, with the tail-end alkyl carbons anchored at z = 0. The ABF method is then used to drive insertion of the phosphate to multiple depths within this flat C16GGGhex layer in order to calculate the associated free energy profile. 53

4.6 Free energy profile for phosphate binding to a flat layer of C16GGGhex. A clear free energy minima is found for low pH conditions, spanning a width of approximately 2 nm; this is in agreement with our single-chain results, which indicated that phosphate is able to bind at low pH in a variety of locations along the GGGhex segment. The high pH free energy profile shows a very small minima of 2.5 kJ/mol, indicating a small amount of stability imparted by the insertion of phosphate into a dense assembly of C16GGGhex; besides this feature, there is no indication that phosphate binding occurs at high pH. 54

5.1 Schematic of order parameters S_T and S_R, which characterize DNA translocation and DNA rotation, respectively. These order parameters are defined in Section 5.2.3. 66

5.2 Schematic of order parameters θ_{forward} and θ_{backward} characterizing DNA breathing, in which strands of DNA spontaneously unwrap and rewrap from the histone complex. Each angle is calculated from two vectors: vector **b** points from the histone complex center of mass to the dyad, and vectors **a** and **c** point along either end of the DNA strand, from the 30th base pair to the first and endmost base pair. 66

5.3 Schematic of the loopiness order parameter, which characterizes the extent to which DNA bulges away from the histone octamer. Calculation of this order parameter is described in Section 5.2.5. 67

5.4 Schematic showing two proposed nucleosome repositioning mechanisms: loop propagation and twist diffusion. The histone complex is represented in red, and DNA in blue. S_T and S_R quantify loop propagation and twist diffusion, respectively; definitions for these order parameters are introduced in the Methods section. Individual repositioning propensities for sequences A, B, and C are also shown. 68

5.5 Eigenvalue spectra for sequences A, B, and C. Note that sequences A and B exhibit hierarchical character, indicated by multiple gaps in the eigenvalue spectra. Both spectra show three dominant eigenvalues, followed by three moderate eigenvalues, suggesting that three major slow dynamical modes dominate the system, while three less significant modes still contribute to the system dynamics. . . . 69

5.6 2- and 3-dimensional diffusion map embeddings of S_T for all sequences. DNA translocation correlates with ψ_2 for sequence A, indicated by the gradient in S_T along the vertical ψ_2 axis. DNA translocation correlates with ψ_1 for sequences B and C, indicated by the gradient in S_T along the horizontal ψ_1 axis. 70

5.7 (a-c): 3-dimensional diffusion map embeddings of S_R, the order parameter that characterizes DNA rotation, for all sequences. For clarity, datapoints with greater values of S_R are shown at higher layers of the plot. There is no correlation of S_R with top non-trivial eigenvectors for sequences A and B; (d) 2-dimensional diffusion map embedding of S_R for sequence C. ψ_1 correlates with cycles of DNA rotation, as indicated by the periodic bands of S_R along ψ_1; (e) effective free energy constructed from the diffusion map embedding for sequence C. Effective free energy is calculated by histogramming datapoints for sequence C in ψ_1 and S_R, normalizing each histogram bin by the total number of datapoints to calculate probabilities, and then taking the negative log of each bin. The resulting density plot exhibits a periodic banded structure reminiscent of the free energy landscape for sequence C constructed by conventional methods by Lequieu et al.,[60] which is plotted in (f) in S_T vs S_R. Note that S_T was previously found to correlate with ψ_1 for sequence C; the diffusion map has effectively unfurled the same previously calculated free energy landscape. 72

5.8 Two-dimensional embeddings of the average of θ_{forward} and θ_{backward}, which characterizes DNA breathing, for all sequences. The average breathing order parameter for sequence A correlates with ψ_2, as indicated by the gradient in the breathing order parameter along the vertical ψ_2 axis; ψ_2 also correlates with the order parameter characterizing DNA translocation, S_T, for sequence A, as seen in Figure 5.6. The average breathing order parameter for sequences B and C correlate with ψ_1, as indicated by the gradient in the breathing order parameter along the horizontal ψ_1 axis; this eigenvector also correlates with S_T for these two sequences, again as seen in Figure 5.6. For sequence C, this eigenvector also correlates with S_R, as seen in Figure 5.7. 74

5.9 Two-dimensional diffusion map embeddings of loopiness for all sequences, plotted by moderate eigenvectors ψ_4, ψ_5 and ψ_6. For sequence A, more loopy configurations are isolated by all three moderate eigenvectors. For sequence B, loopy conformations are only isolated by ψ_6. Loopy configurations are not extracted for sequence C. 75

LIST OF TABLES

5.1 DNA sequences used in this work, along with their binding strengths and sequence names used in the literature. 69

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Professor Juan de Pablo, for all of his mentorship and support throughout my time in his group. Thinking back to when I first started as a graduate student, I feel incredibly blessed that he enthusiastically welcomed me to his group when, having applied to graduate school as an experimentalist, I suddenly decided I wanted to switch to computational work. Juan's enthusiasm, dedication, and passion for science have made a deep impression on me throughout my time in graduate school, and, as I graduate into scary times thanks to a global pandemic, I find that I constantly remind myself of Juan's unwavering optimism and drive in the face of any challenge. I could imagine no better place to have spent my Ph.D. than under Juan's wing.

I am also very thankful to have been able to meet, collaborate with, and be guided by many wonderful people, both at the University of Chicago and places farther away. I am especially grateful for the support of my committee members, Professors Matthew Tirrell and Andrew Ferguson. I am also very thankful towards Professors Paul Nealey, Giulia Galli, Risi Kondor, Moshe Gottlieb, Chong Liu, Alex High, Nicholas Abbott and Jonathan Whitmer, who have all taught me so much, looked out for me, and really broadened my Ph.D. experience. I would like to thank Alejandro Hurtado-Londono and Alex Flage at 3M for many fun and exciting discussions, as well as the opportunity to explore the intersection of my research with industrial applications. Thank you to Professor Teresa Lopez-Leon, who graciously welcomed me to visit her lab in Paris, and to Professors Walter Chapman and Martin Cismoni, who made it possible for me to travel to and present my work at EQUIFASE in Argentina. A big thank you to the Chicago Center for Teaching, especially Kiki Zissimopoulos, for their support during the last year of my Ph.D, as well as the opportunity to hone my teaching skills alongside my research. I would also like to thank Rovana Popoff, Novia Pagone, Diana Morgan, David Taylor, Lisa Abston, Janet Boland, Sandra Marijan, and Heather Crews, who don't get thanked enough for everything they do behind the scenes.

I also want to thank everyone from the de Pablo group, who have been an incredible group of outstanding colleagues and friends. Lucas Antony, Kyle Hoffmann, and Aaron Fluitt helped guide me into the world of protein simulations, and I am especially grateful for their generous pointers and advice, as well as their endless patience and kindness in answering every one of my little questions when I first started my Ph.D. I have been so lucky to work with and learn from Josh Lequieu, whose positive attitude and cheerful curiosity I try to cultivate in myself, as well as Whitney Fowler and Chuting Deng, two of the strongest, most tenacious, and most kind people I have had the pleasure of collaborating with. I don't know how I could have made it through graduate school without my dear friends Cody Bezik, Alec Bowen, Viviana Palacio-Betancur, Josh Moller, and my deskmate and huge pillar of support Emre Sevgen; I will miss our coffee walks and wish we could go on one last one where I could give each of you an infinite cup to thank you for all the support and joy you've given me. I would also like to thank Sumi Hur, Jian Qin, Sanaz Sadati, Yamil Colón, Mike Webb, Nick Jackson, Grant Garner, Hadi Ramezani-Dakhel, Ye Zhou, Weiwei Chu, Marat Andreev, Liza Lee, Jiyuan Li, Boyuan Yu, and Phil Rauscher. A big thank you to Gabi Basel, whose determination, creativity and grit felt so contagious and inspiring that it made me wonder if I was really mentoring her, or if it was actually the other way around.

Finally, I need to thank my friends and family for all of their love and support. I want to thank my parents for always believing in me, for teaching me to be resilient, and for all the opportunities they've given me, and my sister for bringing so much joy into my life. To Julian Grove, thank you for your unwavering support, especially in the final stretch of the Ph.D. To Arin Greenwood, Elizabeth Michiko Ashley, Lily Delalande, Ruben Waldman, Sam Passaglia, and especially Eli Alster, for so much laughter, and for cheering me on and reminding me there's life beyond work. To my Caltech friends, who have propped me up with their endless encouragement despite being flung all over the world after college. Finally, thank you to Professor Julie Kornfield, who helped me see myself as a researcher, and to Dr. Myra Halpin at NCSSM for introducing me to the world of research in the first place.

ABSTRACT

Whether in designing novel materials or simply sustaining basic biological function, the dynamics of biological and bio-inspired macromolecules are key in multiple processes impacting daily life. These dynamics involve a variety of scenarios, including the self-assembly of biomacromolecules, their native dynamics within living cells, and their use in functional materials in order to bind with specific foreign species. In this dissertation, a multitude of tools in the molecular simulation arsenal are deployed to investigate biomacromolecule dynamics in all three scenarios. We begin by using atomistic molecular dynamics to study early-stage aggregation of human islet amyloid polypeptide (hIAPP), an amyloid-forming protein implicated in type II diabetes. By applying the finite temperature string method, we identify potential pathways for the first stages of self-assembly of hIAPP into dimers, as well as relevant aggregation intermediates and their relative stabilities. We then extend our investigation of hIAPP to the formation of trimers, for which we examine multiple possible aggregation mechanisms and study their fundamental mechanistic and thermodynamic differences. We then consider the design of a peptide amphiphile, consisting of a polypeptide chain attached to an alkyl chain that drives self-assembly. We examine the dynamics and energetics of a candidate peptide amphiphile binding to phosphate, which may be harnessed for the sustainable sequestration of phosphate from wastewater. Finally, we proceed to apply a combination of molecular dynamics and nonlinear manifold learning techniques to identify the key dynamical motions of the nucleosome, another biological macromolecular system consisting of 147 base pairs of DNA wrapped around a complex of eight histone proteins, whose sequence-dependent behavior affects critical functions including gene expression and DNA replication.

CHAPTER 1

INTRODUCTION

An incredible amount of natural complexity is found in biological systems, often involving macromolecules such as proteins, DNA, and RNA. The folding, assembling, and binding abilities of these macromolecules underlie their essential roles in the biological processes that sustain life, in contexts ranging from the successful folding of a protein into a structure with specific functionality, to the assembly of multiple proteins into molecular machines, to the compact encoding and storage of genetic information by a combination of proteins and DNA. Studies of these systems not only provide a basis with which to understand the living world around us, but also serve as inspiration for engineering new materials that exhibit the molecular-level specificity, selectivity, and precision often observed in nature.

In this dissertation, we use the tools of molecular simulation and data-driven analysis to study the dynamics of multiple biological macromolecule systems. We examine systems of greater size and complexity as we progress through the dissertation, beginning with the oligomerization of a naturally occurring protein implicated in the onset of disease. We then progress to an engineered self-assembling peptide amphiphile system, designed for the sequestration and recycling of phosphate from wastewater, before proceeding to the study of the dynamics of the nucleosome, a DNA-protein complex responsible for the successful packaging of DNA into chromosomes.

In Chapter 2, we examine the dimerization of human islet amyloid polypeptide (hIAPP or human amylin). Amyloid aggregates of hIAPP have long been implicated in the development of type II diabetes. While hIAPP is known to aggregate into amyloid fibrils, it is the early-stage prefibrillar species that have been proposed to be cytotoxic. A detailed picture of the early-stage aggregation process and relevant intermediates would be valuable in the development of effective therapeutics. We use atomistic molecular dynamics simulations with a combination of enhanced sampling methods to examine the formation of the hIAPP dimer in water. Bias-exchange metadynamics calculations reveal relative conformational stabilities of the hIAPP dimer. Finite temperature string method calculations identify pathways for dimer formation, along with relevant free energy barriers and intermediate structures. We show that the initial stages of dimerization involve crossing a substantial free energy barrier to form an intermediate structure exhibiting transient β-sheet character, before proceeding to form an entropically stabilized dimer structure.

In Chapter 3, we extend our investigation of amylin aggregation to study hIAPP trimerization. We use atomistic molecular dynamics simulations with the finite temperature string method to identify and compare multiple pathways for hIAPP trimer formation in water. We focus on the comparison between trimerization from three disordered hIAPP chains (which we call "3-chain assembly") and trimerization from an hIAPP dimer approached by a single disordered chain (called "2+1 assembly"). We show that trimerization is process uphill in free energy, regardless of the trimerization mechanism, and that a high free energy barrier of 40 k_BT must be crossed in 2+1 assembly compared to a moderate barrier of 12 k_BT for 3-chain assembly. We find this discrepancy to originate from differences in molecular-level water interactions involved in the two trimerization scenarios. Furthermore, we find that the more thermodynamically favorable 3-chain assembly begins from a previously identified dimer intermediate exhibiting transient β-sheet character, which is then incorporated into a similar trimer intermediate, suggesting stepwise aggregation dynamics.

We then proceed to study a system in which naturally occuring peptide sequences are harnessed for the design of functional materials in the context of the recovery of valuable resources from wastewater, which is becoming increasingly important as global population rises and natural resources are depleted. One such resource is phosphate, which is critical for its use in fertilizers in maintaining food production worldwide and lacks any viable substitute. Biologically-inspired peptide amphiphiles are a particular type of material that can address this goal of sequestering phosphate from wastewater, by incorporating a phosphate-binding peptide sequence with an alkyl chain that drives self-assembly to form a self-assembling micellar structure with phosphate-sequestering properties. In Chapter 3, we investigate the preliminary peptide amphiphile candidate C16GGGhex, which is made up of a 16-carbon alkyl tail connected to a biomimetic, pH-responsive, and phosphate-binding hexapeptide via a short peptide linker. We use a combination of molecular dynamics and enhanced sampling methods to study the potential of C16GGGhex for efficient phosphate capture and release at high and low pH conditions. Screening and clustering calculations show that phosphate may bind with C16GGGhex at multiple locations along its peptide region, not solely at the known phosphate-binding hexapeptide. Adaptive biasing force (ABF) simulations of both single C16GGGhex chains and a flat layer of C16GGGhex indicate preferential binding of phosphate at low pH, with three distinct phosphate-binding locations identified in single-chain studies, while no preferential binding is observed at high pH.

In Chapter 5, we expand our work to the study of another essential biomolecule—DNA—while addressing the challenge of identifying effective collective variables in molecular simulations of complex systems. We use a nonlinear manifold learning technique known as the diffusion map to extract key dynamical motions from a complex biomolecular system known as the nucleosome: a DNA-protein complex consisting of a DNA segment wrapped around a disc-shaped group of eight histone proteins. We show that without any *a priori* information, diffusion maps can identify and extract meaningful collective variables that characterize the motion of the nucleosome complex. We find excellent agreement between the collective variables identified by the diffusion map and those obtained manually using a free energy-based analysis. Notably, diffusion maps are shown to also identify subtle features of nucleosome dynamics that did not appear in those manually specified collective variables. For example, diffusion maps identify the importance of looped conformations in which DNA bulges away from the histone complex that are important for the motion of DNA around the nucleosome. This work demonstrates that diffusion maps can be a promising tool for analyzing very large molecular systems and for identifying their characteristic slow modes.

Finally, we conclude with an overview of our results in Chapter 6. We summarize the ways in which this dissertation has demonstrated how modern molecular simulation techniques and data-driven approaches are vital tools for understanding complex systems of biological macromolecules, both in natural and artificial contexts, and provide an outlook on future work.

CHAPTER 2

EARLY-STAGE HUMAN ISLET AMYLOID POLYPEPTIDE AGGREGATION: MECHANISMS BEHIND DIMER FORMATION

Amyloid aggregates of human islet amyloid polypeptide (hIAPP or human amylin) have long been implicated in the development of type II diabetes. While hIAPP is known to aggregate into amyloid fibrils, it is the early-stage prefibrillar species that have been proposed to be cytotoxic. A detailed picture of the early-stage aggregation process and relevant intermediates would be valuable in the development of effective therapeutics. Here, we use atomistic molecular dynamics simulations with a combination of enhanced sampling methods to examine the formation of the hIAPP dimer in water. Bias-exchange metadynamics calculations reveal relative conformational stabilities of the hIAPP dimer. Finite temperature string method calculations identify pathways for dimer formation, along with relevant free energy barriers and intermediate structures. We show that the initial stages of dimerization involve crossing a substantial free energy barrier to form an intermediate structure exhibiting transient β-sheet character, before proceeding to form an entropically stabilized dimer structure. This chapter is reproduced from [38].

2.1 Introduction

Amyloidogenic proteins have long been implicated in a host of human diseases. These proteins exhibit a shared tendency to self-assemble into aggregates known as amyloid, which share morphological properties, including a fibrillar shape with heavily β-sheet secondary structure [70]. One such amyloidogenic protein is human islet amyloid polypeptide (hIAPP or human amylin); this 37-residue polypeptide hormone is co-secreted with insulin in the pancreas by β-cells and plays a role in regulating blood glucose levels [72, 41]. Heavily β-sheet amyloid aggregates of hIAPP have been pathologically linked to the loss of pancreatic

β-cells and the onset of type II diabetes [119], prompting studies of hIAPP fibrils and their formation.

Extensive characterization efforts have probed the structure of the mature hIAPP fibril, including the use of solid-state NMR (ssNMR) experiments to propose the arrangement of individual hIAPP chains within a mature fibril. In this proposed model, hIAPP monomers are stacked along the fibril axis, and each monomer is arranged in a U-shaped conformation consisting of two β-strands (residues 8-17 and 28-37), connected via a loop region. Parallel β-sheets are formed as monomeric hIAPP stack to form the mature fibril structure. This model is consistent with two-dimensional infrared spectroscopy (2D-IR) experiments [117], as well as electron paramagnetic resonance (EPR) measurements [4]. Furthermore, recent experiments and simulations have studied amylin in the presence of inhibitors [6, 122, 83], binding of amylin to metals [121], interactions between amylin and its mutants or other amyloidogenic proteins [48, 47], as well as amylin behavior at a membrane[75, 123, 122, 29, 21, 66] and structural rearrangements during aggregation [104].

Mature fibrils are relatively biologically inert, and the formation of prefibrillar species, or protofilaments, has been linked to cytotoxicity [103, 50, 18, 82], prompting a shift toward the study of early-stage amyloid aggregates. Dimers, trimers, and larger oligomers have been proposed to be responsible for disrupting cell membranes, inhibiting metabolic functions, inducing oxidative stress, and triggering apoptosis [50, 69, 117, 78, 91]. Experimental evidence includes observations of membrane leakage prior to mature fibril formation [100] and of disrupted cell membranes isolated from areas of fibril growth [13]. In order to better understand the role of hIAPP in disease onset, as well as to design effective therapeutics, it is crucial to uncover the mechanisms of early-stage fibril formation, as well as identify any relevant intermediate structures in the aggregation process.

In particular, residues 20-29 in hIAPP are suspected to play a key role in amyloid formation [120]. This 10-residue segment is itself amyloidogenic, and mutations to this sequence suppress amylin aggregation [120, 79, 92, 1]. However, residues 20-29 are not located within the heavily β-sheet fibril core in the ssNMR-based fibril model [70]. 2D-IR experiments and molecular dynamics (MD) simulations have previously been combined to identify a key intermediate in the early-stage aggregation of hIAPP, which featured the FGAIL region in residues 23-27 in a transient parallel β-sheet prior to the formation of the final U-shaped configuration [14]. In addition to this FGAIL region, residues L12A13 have recently been suggested to form a stacked turn or disordered β-sheet during the aggregation through 2D-IR experiments with dihedral indexing [73]. Furthermore, Chiu and de Pablo have utilized MD simulations combined with bias-exchange metadynamics to study the mechanism by which two disordered hIAPP monomers assemble into a U-shaped dimer [14, 20], and more recent work has extended this approach to investigate dimer formation in the presence of a lipid membrane [66]. Two potential dimerization mechanisms were proposed, both of which exhibit intermediate parallel β-sheet structure in residues 23-27 (FGAIL region). The disordered dimer was found to be less thermodynamically stable than both the intermediate and the final U-shaped dimer; however, the work concludes with a cautionary note that these results are highly dependent on the force field used (GROMOS96 53a6) [86, 20]. Subsequent force field reviews suggested that GROMOS96 53a6 tends to over-stabilize the formation of β-sheet in both human and rat islet amyloid polypeptide (rIAPP) [46], as well as the amyloidogenic polypeptide polyglutamine [32]. Newer force fields were shown to more accurately capture experimentally determined properties, for example C_α secondary NMR chemical shifts: one such force field is AMBERff99SB*-ILDN [65, 64].

Although GROMOS96 53a6 does predict an intermediate structure consistent with previous experimental and computational studies, given the findings of Hoffmann et al. [46], further investigation of amylin aggregation warrants the use of a force field that better captures the conformational behavior of the hIAPP molecule, especially if studies are to be extended to higher order aggregates. In this work, we seek a more accurate understanding of early-stage hIAPP aggregation in water using the AMBERff99SB*-ILDN force field. We use bias-exchange metadynamics to reveal the free energy landscape for hIAPP dimerization under the new, more accurate force field. Furthermore, we go beyond past studies of the dimer and employ finite temperature string method calculations to unveil specific aggregation mechanisms for the hIAPP dimer, as well as the associated changes in free energy, revealing a substantial free energy barrier associated with the formation of an intermediate β-sheet structure prior to formation of a locally stable structured amylin dimer.

2.2 Results and Discussion

2.2.1 Free Energy Landscape for the hIAPP Dimer

Bias-exchange metadynamics (as described in Methods) using the AMBERff99SB*-ILDN force field was used to produce the free energy landscape of the hIAPP dimer in solution (Figure 2.1a), plotted as a function of two collective variables (CVs), $Q_{\text{res 8-16}}$ and $Q_{\text{res 27-35}}$. The two Q parameters measure similarity of residues 8-16 and 27-35 in the simulated structure to the ideal U-shaped dimer, which is extracted from the ssNMR mature fibril structure. Further details on the chosen collective variables are found in Methods. The global free energy minimum is found at $(Q_{\text{res 8-16}}, Q_{\text{res 27-35}}) = (0.43, 0.38)$. Two local minima corresponding to fully dimerized structures are found at the upper right of the plot at $(0.88, 0.88)$ and $(0.88, 0.68)$, with free energies of 7.4 kJ/mol and 6.8 kJ/mol respectively. Additional local minima are found at $(0.83, 0.33)$, with a free energy of 6.5 kJ/mol, and $(0.23, 0.68)$, with a free energy of 7.3 kJ/mol; these minima correspond to partially dimerized structures.

Several features of the free energy landscape can be pieced together to understand the dimerization process. First, the global minimum is found in a single, wide basin in the lower left quadrant of the free energy landscape. In Figure 2.1a, conformations within 5 kJ/mol of the global minimum are bounded by the second contour line out from the global minimum, shown in yellow. Structures in this lower left quadrant correspond to low values of $Q_{\text{res 8-16}}$ and $Q_{\text{res 27-35}}$; these conformations show little similarity to the ideal U-shaped dimer. The free energy minimum that most closely matches the U-shaped dimer is

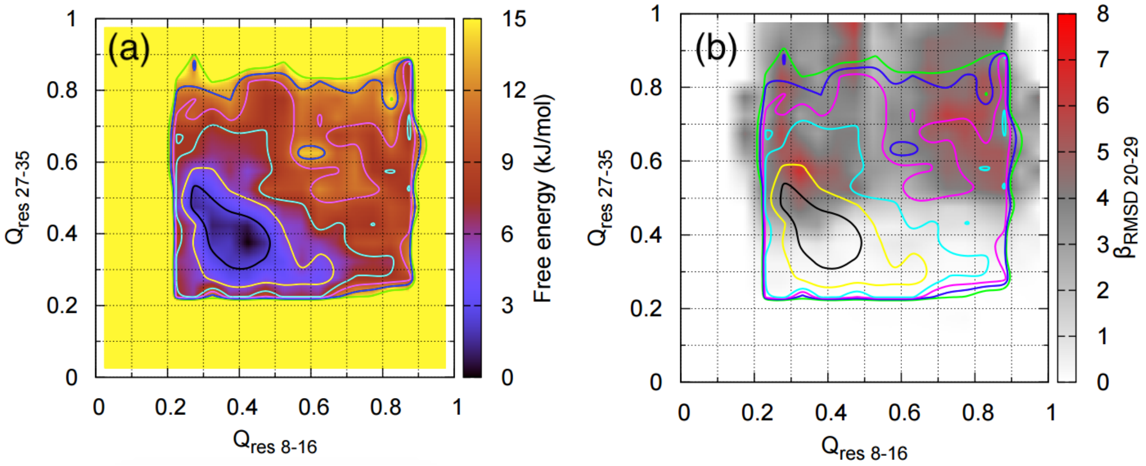


Figure 2.1: (a) Free energy landscape of the hIAPP dimer as a function of $Q_{\text{res 8-16}}$ and $Q_{\text{res 27-35}}$, obtained via BE-MetaD and the AMBERff99SB*-ILDN force field. Free energy values are relative to the global minimum at $(Q_{\text{res 8-16}}, Q_{\text{res 27-35}}) = (0.43, 0.38)$. Contour lines are plotted for free energies from 2.5 kJ/mol (black) to 15 kJ/mol (green), with stride 2.5 kJ/mol. All free energies greater than or equal to 15 kJ/mol are plotted in yellow. Two local minima corresponding to dimer structures are located, at $(0.88, 0.88)$ and $(0.88, 0.68)$. (b) Average amount of parallel β-sheet spanning residues 20-29 in the hIAPP dimer in solution, as measured by $\beta^{\text{parallel}}_{\text{RMSD 20-29}}$, as function of $Q_{\text{res 8-16}}$ and $Q_{\text{res 27-35}}$. The value of $\beta^{\text{parallel}}_{\text{RMSD 20-29}}$ ranges from 0 (no parallel β-sheet) to 8 (completely parallel β-sheet). Statistics were collected from the full production simulation time of all six BE-MetaD replicas. Free energy contours from (a) are superimposed on the plot.

found in the upper right quadrant at $(0.88, 0.88)$, elevated over the global minimum by 7.4 kJ/mol, or approximately 3 k_BT at room temperature. The relative depths, widths, and locations in CV space of these two minima suggest that the conformational ensemble of the hIAPP dimer is largely dominated by disordered structures dissimilar from the U-shaped dimer. Additionally, the free energy minima in the upper right quadrant at $(0.88, 0.88)$ and $(0.88, 0.68)$ are more elongated along the $Q_{\text{res 27-35}}$ axis versus the $Q_{\text{res 8-16}}$ axis, suggesting that C-termini β-strands are more flexible than the N-termini β-strands. Furthermore, the conformational terrain is relatively smooth outside of the global minimum, with no local free energy minima deeper than 2.5 kJ/mol below their immediate surroundings; the major barrier to the dimerization process is escaping the wide well surrounding the global minimum centered at $(0.43, 0.38)$.

A number of qualitative differences arise between the AMBERff99SB*-ILDN and GROMOS96 53a6 models. The GROMOS model previously suggested that the U-shaped dimer is found at the global free energy minimum, at 1.1 kJ/mol more stable than the disordered dimer. The GROMOS model also identified multiple metastable states, with free energies within k_BT of the global minimum; each of these metastable states featured a significant amount of β-sheet structure. In comparison, the free energy landscape obtained using AMBERff99SB*-ILDN features a U-shaped dimer at 7.4 kJ/mol above the disordered dimer in free energy, with a single wide basin centered around the global minimum containing disordered conformations. This discrepancy is consistent with the findings of Hoffmann et al. that found that GROMOS96 53a6 overstabilizes β-sheet secondary structure in hIAPP and that the AMBER model can more accurately reproduce the true structure of hIAPP in solution [46]; this prompts reexamination of the true mechanistic details behind the hIAPP dimerization process.

While Figure 2.1a highlights individual conformational clusters and their relative stabilities, little mechanistic information can be drawn from the free energy surface. In order to extract information regarding the role of a possible transient β-sheet intermediate formed in residues 20-29, we overlay the average amount of parallel β-sheet in residues 20-29 onto the free energy contours from Figure 2.1a. The overlay pinpoints two clusters of parallel β- sheet character in residues 20-29: one tight cluster at $(0.33, 0.58)$ approximately 5.0 kJ/mol above the global minimum, and a second diffuse cluster at $(0.73, 0.73)$ approximately 10 kJ/mol above the global minimum. If a β-sheet intermediate is indeed formed in residues 20-29 during the dimerization process, the transition pathway must pass through these regions of elevated parallel β-sheet mapped in Figure 2.1b.

However, there are two such clusters of high β-sheet character in residues 20-29 and little information to distinguish whether these two separate regions in CV space share degenerate hIAPP conformations. It is difficult to pinpoint a specific aggregation pathway on this free energy landscape, even moreso considering that the two Q parameters are only dependent on a subset of all the residues in the system. In order to obtain details on the mechanism of aggregation, we perform finite temperature string method calculations using collective variables that may better capture conformational changes across the entire polypeptide. While performing yet another round of metadynamics with these more general collective variables might be prohibitively expensive, string methods inherently focus on the transition region, making them an attractive alternative for studying the dimerization process.

We use the finite temperature string method to examine the hIAPP dimerization process along two collective variables: the amount of parallel β-sheet formed between hIAPP chains and the distance between the centers of mass of each hIAPP monomer.

2.2.2 hIAPP Dimerization Pathway

Figure 2.2 shows the final transition pathway for the hIAPP dimerization process, obtained from finite temperature string method calculations (see Methods). One end of the string corresponds to the disordered state of the hIAPP dimer (labeled Structure I, located at $d_{\text{COM}} = 1.31$ nm and $\beta_{\text{RMSD}}^{\text{parallel}} = 4.32$) and the opposite end of the string corresponds to the fully formed dimer (Structure V, at $d_{\text{COM}} = 0.40$ nm and $\beta_{\text{RMSD}}^{\text{parallel}} = 24.80$). Representative snapshots of the system are shown alongside the pathway in Figure 2.2, highlighting key structural changes that occur during the dimerization process.

In the disordered configuration shown in Panel I, no secondary structure is formed between the two amylin chains. However, any individual disordered chain may independently form α-helix or β-strand motifs. Panel II highlights an intermediate structure, consistent with previous studies, exhibiting parallel β-sheet structure that has been suspected to play a role in amyloid aggregation [120]. Increasing amounts of parallel β-sheet is formed between the two amylin monomers as dimerization progresses, as witnessed in Panels III and IV, which show parallel β-sheet forming in the C-termini before advancing to the N-termini to form the full fibrillar dimer in Panel V.

The free energy profile along this dimerization pathway is shown in Figure 2.3. An

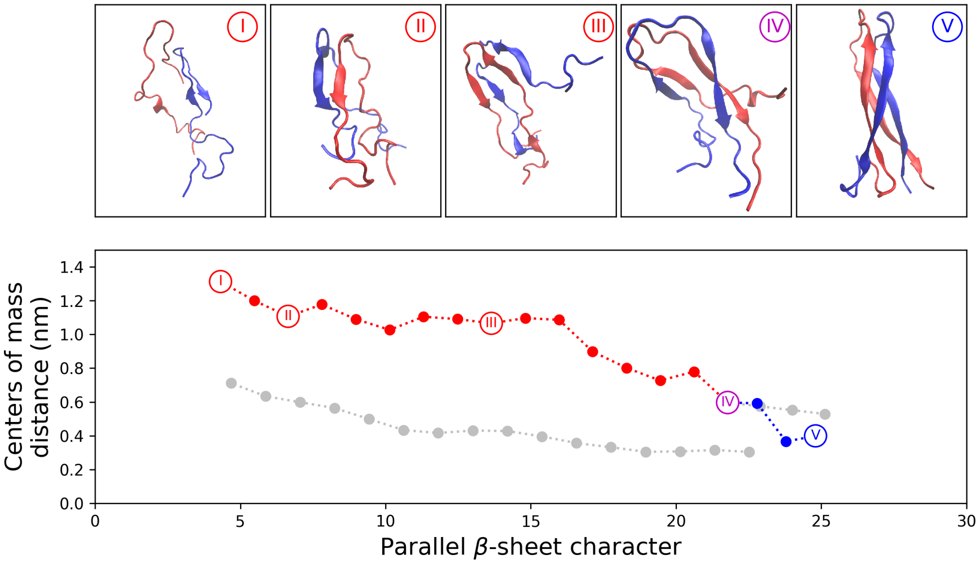


Figure 2.2: Dimerization pathway obtained via finite temperature string method. Five representative snapshots illustrate conformational changes that take place during dimer formation. Water and counterions are not shown for clarity. The 35-node pathway is composed of the results of a 32-node and smaller 4-node string method, as described in Methods. Grey points show initial configurations for the 32-node and 4-node string method calculations.

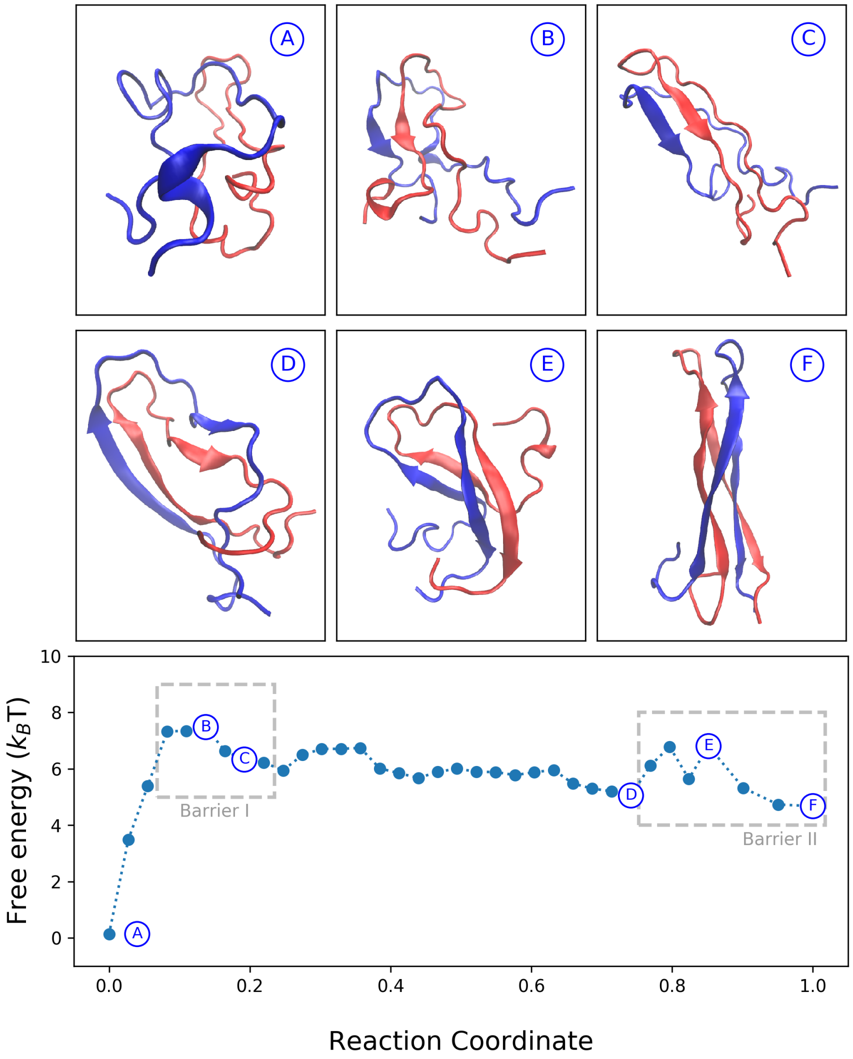


Figure 2.3: Free energy profile along the dimerizaton pathway obtained via finite temperature string method. The reaction coordinate proceeds from 0.0 (disordered state) to 1.0 (fully formed dimer). Free energy is calculated as described in Methods, sampling 35 Voronoi cells for 150 ns each. Representative snapshots illustrate key conformational changes, including the formation of a transient β-sheet structure after surpassing the initial free energy barrier of approximately 7 k_BT. Average secondary structure per residue across the entire dimerization pathway is presented in Figure 2.4.

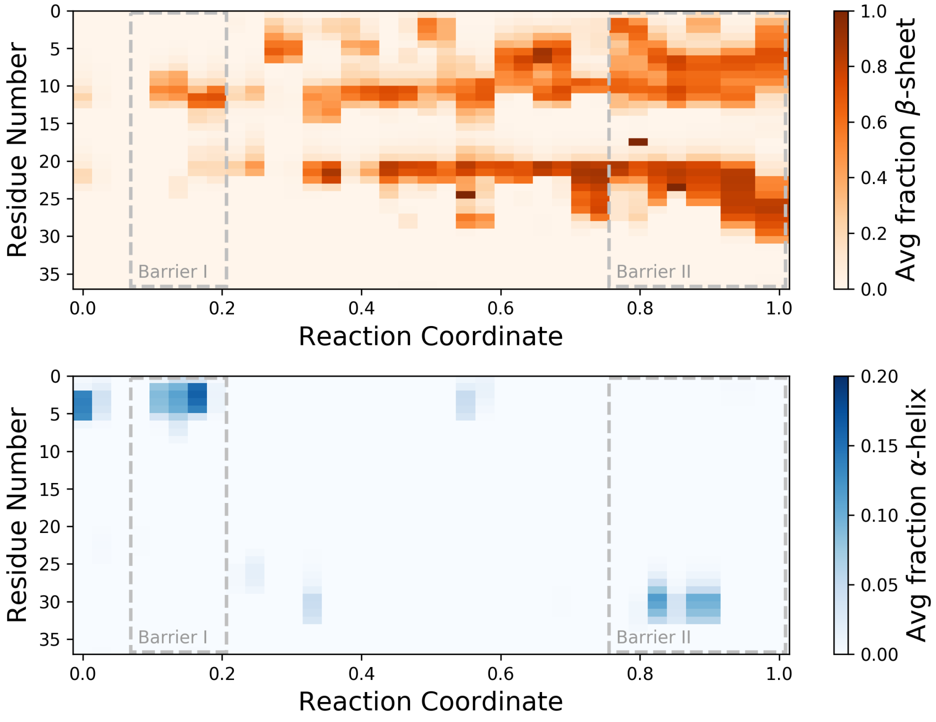


Figure 2.4: Average β-sheet and α-helix secondary structure at each of the points where a free energy calculation was performed in Figure 2.3, plotted by residue (ranging from 1 to 37) versus reaction coordinate (from the disordered state at 0.0 to fully dimerized state at 1.0). Secondary structure was assigned using the DSSP algorithm [53] and averaged over all 150 ns of sampling for each bin. The Barrier I and II regions corresponding to the barriers highlighted in Figure 2.3 are marked in both plots. As Barrier I is crossed, a transient β-sheet is formed in residues 19-24 and in the L12A13 region. The crossing of Barrier II corresponds to simultaneous loss of β-sheet in the C-termini and increased β-sheet in the N-termini, with regions of high β-sheet character gradually moving out towards the termini as the full dimer is formed, accompanied with a drop in free energy.

initial free energy barrier of approximately 7 k_BT is discovered, and the images provided in Figure 2.3 indicate that this barrier corresponds to the formation of the intermediate β-sheet structure. As the system climbs the 7 k_BT energy barrier, the two amylin monomers approach each other and align; as the system traverses past the peak of the barrier, the transient intermediate β-sheet forms, accompanied by a drop in free energy.

Figure 2.4 shows the average secondary structure (calculated via the DSSP algorithm [53]) per residue for the dimerization process, with barriers corresponding to those in Figure 2.3 marked. The intermediate β-sheet structure formed in the Barrier I region is found to be localized not only in the previously proposed region in residues 20-29 but also the L12A13 region recently proposed by Maj and coworkers [73]. Note that this is slightly shifted toward the N-termini compared with the GROMOS model.

A small free energy barrier is observed at the opposite end of the pathway, on the order of 2 k_BT. By examining the corresponding snapshots, it becomes clear that this barrier is associated with a conformational rearrangement of the two amylin monomers, transitioning from two extended chains stacked side-by-side in parallel to a more compact "ribbon-like" structure exhibiting a slight twist. This rearrangement is coupled with additional formation of N-termini β-sheet structure, as reflected in the plot of secondary structures in Figure 2.4. Figure 2.4 also reveals that the additional formation of β-sheet in the N-termini, and its associated free energy stabilization, comes at the cost of relinquishing a fraction of β-sheet in the C-termini. Furthermore, formation of the fully dimerized structure corresponds to a shift in localization of β-sheet character away from the center of the hIAPP chain towards the termini.

Additional mechanistic details can be uncovered by tracking the entropic and enthalpic contributions to the changes in free energy (Figure 2.5). This is accomplished by calculating change in potential energy ΔU and change in free energy ΔA for each sampled Voronoi cell along the reaction coordinate with respect to the disordered bin, followed by calculating the entropic changes using $\Delta A = \Delta U - \Delta TS$. The enthalpic and entropic profiles are shown in Figure 2.5. Enthalpic and entropic changes fluctuate as dimerization progresses, before a sharp jump in entropic contributions occurs, corresponding to Panel E in Figure 2.3. By overlaying multiple RMSD-aligned snapshots taken from the same trajectory as Panel E (Figure 2.6), we see the jump in entropy reflected in the conformational degeneracy of the dimer at both termini; while the core of the dimer remains compact, the ends of each chain explore many configurations. The entropic contribution then drops as we move to Panel F in Figure 2.3, stabilizing the final dimer structure by approximately 2 k_BT.

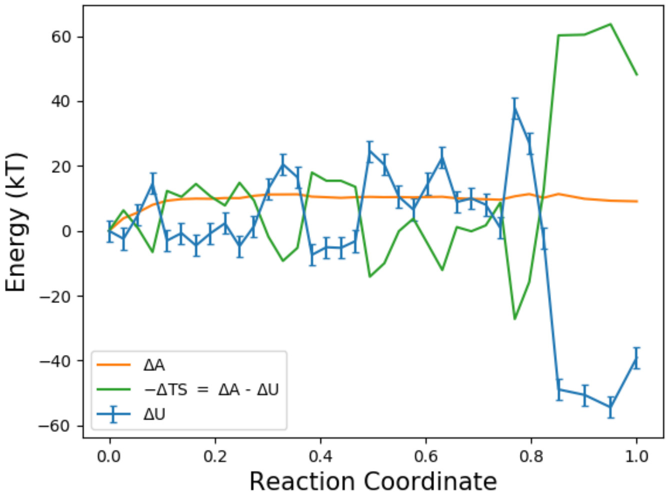


Figure 2.5: Free energy decomposition into enthalpic (ΔU) and entropic ($-\Delta TS$) contributions for dimerization. ΔA is taken from the free energy calculation, performed as described in Methods. Average potential energy is calculated from each of the 35 Voronoi cells sampled during free energy calculation, from a total of 15003 snapshots per cell. Entropic contributions are then calculated as $-\Delta TS = \Delta A - \Delta U$.

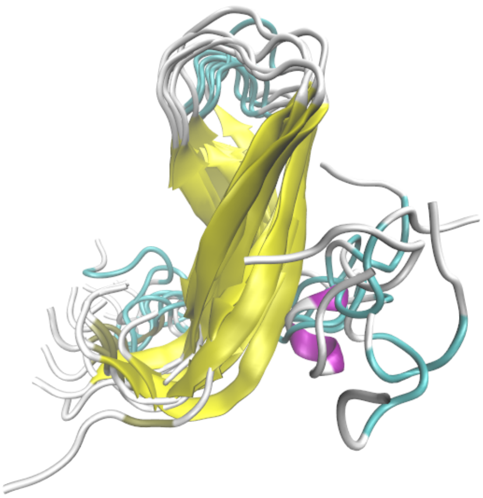


Figure 2.6: Six RMSD-aligned snapshots of the dimer, taken from the trajectory corresponding to the free energy barrier marked by snapshot E in Figure 2. While the core of the dimer structure remains compact across the multiple snapshots, the termini of each hIAPP molecule explore more freely, contributing to the jump in entropy found in Figure 2.5. Structures were aligned using VMD [9, 49].

2.3 Conclusions

Bias-exchange metadynamics simulations have been used to study the thermodynamics of hIAPP dimerization using the AMBERff99SB*-ILDN force field. A global free energy minimum corresponding to the disordered dimer has been identified, as well as a metastable free energy minimum corresponding to the fully-formed dimer, whose value is 7.4 kJ/mol or approximately 3 k_BT higher.

The dimerization pathway for hIAPP, determined from finite temperature string method calculations, reveals an energy barrier of approximately 7 k_BT associated with the formation of an intermediate β-sheet structure. Interestingly, this structure is localized in the previously proposed region in residues 20-29, as well as in the more recently proposed L12A13 region. Consistent with the results of metadynamics simulations, the fully formed dimer corresponds to a local free energy minimum whose energy is approximately 4.5 k_BT higher than that of the disordered dimer. The fully formed dimer lies in a shallow well of approximately 2 k_BT; importantly, further free energy decomposition suggests that the final dimer structure is entropically stabilized.

While a moderate (7 k_BT) free energy barrier is associated with the formation of the intermediate dimer structure, the question that now arises is whether aggregation will remain uphill in free energy as the oligomer grows larger, or whether the aggregation process will become favorable after a certain size oligomer has been formed, thereby initiating exponential fibril growth. Other remaining questions include whether higher order aggregates nucleate from the globally stable disordered state, an intermediate structure formed during dimerization, or the locally stabilized structured dimer, and how these higher order aggregation events proceed. Such simulations are considerably larger in magnitude and are being pursued in our laboratory – the results will be presented in a future publication. Further investigations of that nature may shed light on the formation of higher order aggregates and the effects of different physiologically relevant environments on the aggregation process.

2.4 Methods

2.4.1 Human Amylin Dimer

We base the design of our dimer system on that of Chiu and de Pablo [20], with the major difference being the change in force field. The simulated system consists of two hIAPP molecules, 26,626 water molecules, and six chloride ions. The amino acid sequence for hIAPP is KCNTATCATQRLANFLVHSSNNFGAILSSTNVGSNTY. The C-termini of each polypeptide is amidated, and the side chains of Cys2 and Cys7 are linked by a disulfide bond. Protonation states of all ionizable functional groups are assigned on the basis of their pKa values in aqueous solution at a pH of 7.0. Each hIAPP molecule carries a formal charge of +3, and chloride counterions are included to ensure zero net charge in the system. Polypeptides and ions were modeled by the AMBERff99SB*-ILDN force field [64, 88], and water was modeled by the TIP3P model [52]. The system was placed in a periodic cubic box of side length 9.4 nm. Coulombic forces were calculated using the particle mesh Ewald algorithm [25, 31], temperature coupling at 298 K was achieved using the Nosé-Hoover thermostat [84], volume was held constant, and simulations were performed with a timestep of 2 fs. The LINCS algorithm was used to constrain hydrogen bond lengths to equilibrium values [43].

2.4.2 Bias-Exchange Metadynamics

Bias-exchange metadynamics (BE-MetaD) simulations were performed to construct a free energy landscape for the hIAPP dimer. BE-MetaD performs conventional metadynamics in parallel on multiple replicas of the system, with atomic coordinates periodically exchanged between replicas à la parallel tempering [89]. As in standard metadynamics [59, 16], the final bias potential consists of a sum of small, Gaussian potentials deposited periodically along the system trajectory in collective variable (CV) space. Free energies are calculated from the combined statistics using the weighted histogram analysis method (WHAM) [58]. In this study, and in the following equations, each replica is biased along one CV at most.

The cumulative bias potential of replica i at time t, denoted V_G^i, is [59]:

$$V_G^i[\xi_i(\mathbf{x}_i(t)), t] = W \sum_{t' \leq t} \exp\left(-\frac{[\xi_i(\mathbf{x}_i(t)) - \xi_i(\mathbf{x}_i(t'))]^2}{2\sigma_i^2} \right) \tag{2.1}$$

where ξ_i denotes the CV, $\xi(t)$ is the atomic coordinate vector, and the times at which Gaussian potentials were previously deposited are t_0. The amplitude and width of the Gaussian potentials are W and σ_i, respectively. Exchanges of the atomic coordinates and velocities are attempted every 50 ps between randomly selected pairs of replicas. Coordinates of replicas i and j are exchanged with probability p_{ij}:

$$p_{ij} = \min\left[1, \exp\left(\frac{V_G^i[\xi_i(\mathbf{x}_i(t)), t] + V_G^j[\xi_j(\mathbf{x}_j(t)), t)] - V_G^i[\xi_i(\mathbf{x}_j(t)), t] - V_G^j[\xi_j(\mathbf{x}_i(t)), t]}{k_B T} \right)\right] \tag{2.2}$$

where k_B is the Boltzmann constant and T is the temperature. Since all replicas are held at the same, constant temperature, the exchange probability is independent of the energy from the true, non-biasing MD interactions. Beyond some elapsed simulation time t_{fill}, the system diffuses freely in CV space, and a time average of the inverse cumulative bias potential is an estimator of the underlying free energy surface. Using WHAM, statistics from all replicas may be combined and the free energy A may be calculated as a function of any CV ξ' [58, 9]:

$$A(\xi') = -k_B T \ln \left(\frac{\sum_i^k n_{\xi'}^i}{\sum_j^k \exp\left[\frac{1}{k_B T} \left(f^j - \overline{V_G^j(\xi')} \right) \right]} \right) \tag{2.3}$$

where k is the number of replicas, and $\overline{V_G^j(\xi')}$ is the average biasing potential acting along ξ' in replica j, and the f^j is the normalization constant for replica j, calculated iteratively through the WHAM algorithm. $n_{\xi'}^i$ represents the number of times that replica i visits CV value ξ'.

Two types of CVs were chosen for the BE-MetaD calculations: $Q_{\text{res u-v}}$, a measure of similarity of to the proposed amylin fibril structure derived from ssNMR experiments, and $\beta^{\text{parallel}}_{\text{RMSD u-v}}$, a measure of parallel β-strand content. In each case, a range of amino acid sequences is bounded by residues u and v. Specifically, $Q_{\text{res u-v}}$ is the RMSD structural similarity of resiudes u through v to the same set of residues from the ssNMR structure [111, 20],

$$Q_{\text{res u-v}} = \left\langle \exp\left[\frac{-(r^{\text{ref}}_{ij} - r_{ij})^2}{9\text{Å}^2}\right]\right\rangle_{i \neq j} \tag{2.4}$$

where r_{ij} and r^{ref}_{ij} are the distances, in Å, between backbone atoms i and j in the sampled and reference configuration, respectively. The angle brackets indicate an average over all pairs of backbone atoms belonging to residues in the sequence bounded by residues u and v. The value of $Q_{\text{u-v}}$ ranges from 0, which indicates no similarity, to 1, which indicates an identical conformation. In this study, we adopt $Q_{\text{res 8-16}}$ and $Q_{\text{res 27-35}}$ as CVs. The reference structures are residues 8-16 and residues 27-35, respectively, of the ssNMR structure [70], i.e. the N- and C-terminal β-strands in the fibril.

The parallel β-sheet content of a particular sequence bounded by residues u and v is defined as [90]:

$$\beta^{\text{parallel}}_{\text{RMSD u-v}} = \sum_{\beta} \frac{1 - \left(\frac{\text{RMSD}}{0.8\text{Å}}\right)^8}{1 - \left(\frac{\text{RMSD}}{0.8\text{Å}}\right)^{12}} \tag{2.5}$$

where the sum runs over all pairs of three-residue segments bounded by residues u and v in both molecules. RMSD measures the root mean square deviation, in Å, of the positions of the N, C_α, C_β, C, and O backbone atoms in those (3 + 3)-residue blocks from those in an ideal, parallel β-sheet. Put another way, $\beta^{\text{parallel}}_{\text{RMSD u-v}}$ counts the number of pairs of three-residue segments similar to the ideal parallel β-sheet, scaled by a switching function.

We adopt $\beta^{\text{parallel}}_{\text{RMSD 8-16}}$, $\beta^{\text{parallel}}_{\text{RMSD 27-35}}$, and $\beta^{\text{parallel}}_{\text{RMSD 20-29}}$ as CVs. Biasing along $\beta^{\text{parallel}}_{\text{RMSD 8-16}}$ and $\beta^{\text{parallel}}_{\text{RMSD 27-35}}$ accelerates sampling of all parallel β-sheets involving those segments, not just the ones in the ssNMR structure. Biasing along $\beta^{\text{parallel}}_{\text{RMSD 20-29}}$ accelerates sampling of parallel β-sheets in the central regions of the molecules, which have been identified as a key intermediate in the formation of fibrils [120].

BE-MetaD simulations were carried out for a system consisting of six replicas. Five replicas were subject to the metadynamics biasing potential acting along one of the five CVs defined above: $Q_{\text{res 8-16}}$, $Q_{\text{res 27-35}}$, $\beta^{\text{parallel}}_{\text{RMSD 8-16}}$, $\beta^{\text{parallel}}_{\text{RMSD 27-35}}$, or $\beta^{\text{parallel}}_{\text{RMSD 20-29}}$. The sixth replica evolved with zero biasing potential, but was allowed to exchange coordinates with the other replicas according to Equation 2.2. For all replicas, $W = 2.0\text{kJ mol}^{-1}$ and Gaussian potentials were deposited every 5 ps. Widths of the deposited Gaussian potentials were $\sigma_i = 0.02$ for $Q_{\text{res 8-16}}$ and $Q_{\text{res 27-35}}$; widths of the deposited Gaussians were $\sigma_i = 0.2$ for $\beta^{\text{parallel}}_{\text{RMSD 8-16}}$, $\beta^{\text{parallel}}_{\text{RMSD 27-35}}$, and $\beta^{\text{parallel}}_{\text{RMSD 20-29}}$. The total simulation time was 500 ns. The filling time t_{fill} was 400 ns. Simulations were conducted using the GROMACS 4.5.5 simulation package [5, 44] and a modified version of the PLUMED 1.3 plugin [12]. WHAM calculations were conducted using the METAGUI plugin for VMD [9, 49].

2.4.3 Finite Temperature String Method

In order to identify a pathway for hIAPP dimerization, we utilize the finite-temperature string method [115], which calculates a transition pathway as a series of local points (or "nodes") connected by a smooth curve (or "string") in collective variable space. Here, we investigate the free energy landscape described in terms of two intuitive collective variables that depend on all residues in the system: (1) parallel beta sheet character $\beta^{\text{parallel}}_{\text{RMSD}}$, following Equation 3.1 and using all residues in the system; and (2) a measure of the spatial separation between individual hIAPP chains. We use the distance between the centers of mass for each hIAPP monomer.

The string is discretized into 16 nodes, located in collective variable space at $\mathbf{z}_\alpha$, where α indicates index along the string ($\alpha = 0, ..., 15$). The nodes split collective variable space into a Voronoi tessellation, where each node has an associated Voronoi cell consisting of the points in CV space closer to itself than any other node. We assume Euclidean geometry. At every string method iteration, each Voronoi cell is sampled such that no bias is applied while the system is within the boundaries of its own Voronoi cell, and a harmonic restraining potential is applied when the system departs from its own cell:

$$V_{\text{Voronoi}} = \begin{cases} 0 & \text{system in cell} \\ k_i(\|\mathbf{z}(\mathbf{x}(t)) - \mathbf{z}_\beta\|)^4 & \text{system out of cell, in cell } \beta \end{cases} \tag{2.6}$$

Each string method iteration samples every Voronoi cell for 100 ps. We track the running average of each node's location in collective variable space since the first iteration $\overline{\mathbf{z}_\alpha}$. At the nth iteration, the string is updated according to:

$$\mathbf{z}_\alpha^{n+1} = \mathbf{z}_\alpha^n - \Delta\tau(\mathbf{z}_\alpha^n - \overline{\mathbf{z}_\alpha}) + \mathbf{r}_\alpha \tag{2.7}$$

where $\Delta\tau$ is chosen to be 0.1, and the smoothing parameter $\mathbf{r}_\alpha$ is equal to 0 for nodes on either end of the string ($\alpha = 0$ or 15), otherwise:

$$\mathbf{r}_\alpha = \kappa N^2 \Delta\tau(\mathbf{z}_{\alpha+1} + \mathbf{z}_{\alpha-1} - 2\mathbf{z}_\alpha) \tag{2.8}$$

where smoothing parameter κ is 0.1, and the number of nodes along the string N is 16. After each update to the string, a cubic spline interpolation is drawn through the 16 nodes, and the nodes are redistributed along the string to maintain equal arc-length between adjacent nodes.

We iterate through these steps until the string converges, after which we run a secondary string method calculation with only $N = 4$ nodes. This additional string method calculation aims to ensure discovery of the true fibrillar state, which BE-MetaD suggests may lie in a relatively narrow, isolated free energy basin. This "miniature" string is initialized from an idealized dimer structure (extracted from the ssNMR mature fibril structure [70]) to the fibrillar end of the original, converged 16-node string, which is held pinned in CV space throughout the evolution of the 4-node string.

Upon convergence of the secondary string, the free energy is computed along the final composite string, consisting of the 16-node string and 4-node string stitched together. This is done by calculating π_α, the equilibrium probability of the system to be found in Voronoi cell α, which is then used to calculate the corresponding free energy A_α [74]:

$$A_\alpha = \frac{1}{k_B T}\log(\pi_\alpha) \tag{2.9}$$

To improve resolution of the resulting free energy profile, we further discretize the original 16 node string to 32 nodes, resulting in the complete pathway being described by a total of $N = 35$ Voronoi cells. Each of the Voronoi cells is sampled using the same soft wall restraints described in Equation 3.2, for 50 ns for multiple runs. For each system sampling cell α, we collect T_α, the total simulation time spent within cell α, as well as $N_{\alpha\gamma}$, the number of times the system escapes into a neighboring cell γ. The equilibrium probabilities π_α are calculated with the following system of equations, where $\nu_{\alpha\gamma} = \dfrac{N_{\alpha\gamma}}{T_\alpha}$ is the rate of escape from cell α into γ:

$$\sum_{\gamma=1}^{N} \pi_\gamma \nu_{\gamma\alpha} = \sum_{\gamma=1}^{N} \pi_\alpha \nu_{\alpha\gamma} \tag{2.10}$$

$$\sum_{\alpha=1}^{N} \pi_\alpha = 1 \tag{2.11}$$

String method simulations were performed using the GROMACS 4.6.7 simulation package [5, 44], the PLUMED 2.1 plugin [12], along with custom code to perform string method calculations.

CHAPTER 3

MOLECULAR INSIGHTS INTO THE ROLE OF WATER IN EARLY-STAGE HUMAN AMYLIN AGGREGATION

Human islet amyloid polypeptide (hIAPP or human amylin) is known to aggregate into amyloid fibrils and is implicated in the development of type II diabetes. Prefibrillar species in particular have been linked to cell loss, prompting detailed investigation of early-stage hIAPP aggregation. Insights into the mechanisms underlying early-stage aggregation and the key intermediate structures formed during aggregation are valuable in understanding disease onset at the molecular level and guiding design of effective therapeutic strategies. Here, we use atomistic molecular dynamics simulations with the finite temperature string method to identify and compare multiple pathways for hIAPP trimer formation in water. We focus on the comparison between trimerization from three disordered hIAPP chains (which we call "3-chain assembly") and trimerization from an hIAPP dimer approached by a single disordered chain (called "2+1 assembly"). We show that trimerization is a process uphill in free energy, regardless of the trimerization mechanism, and that a high free energy barrier of 40 $k_B T$ must be crossed in 2+1 assembly compared to a moderate barrier of 12 $k_B T$ for 3-chain assembly. We find this discrepancy to originate from differences in molecular-level water interactions involved in the two trimerization scenarios. Furthermore, we find that the more thermodynamically favorable 3-chain assembly begins from a previously identified dimer intermediate exhibiting transient β-sheet character, which is then incorporated into a similar trimer intermediate, suggesting stepwise aggregation dynamics. Some background information from Chapter 2 is revisited for thoroughness.

3.1 Introduction

Abnormal aggregation of amyloidogenic proteins is implicated in numerous human diseases, including type II diabetes and various neurodegenerative diseases, such as Alzheimer's disease. In each of these diseases, a particular protein self-assembles into a type of heavily β-sheet fibrillar aggregate known as amyloid.[70] Human islet amyloid polypeptide (hIAPP or human amylin) is one such amyloidogenic protein; this 37-residue hormone is secreted with insulin in the pancreas and is involved in blood glucose regulation.[72, 41]. Formation of amyloid aggregates of hIAPP has been linked to the development of type II diabetes as well as the loss of pancreatic β-cells, [119] which has motivated the study of hIAPP aggregates and the mechanism through which they are formed.

The structure of the mature hIAPP fibril has been studied extensively via various structural characterization methods, including solid-state NMR (ssNMR) experiments, used to identify the how individual hIAPP monomers are arranged within a mature fibril. In this ss-NMR model, hIAPP monomers are stacked one by one in the direction of the fibril axis, with each individual hIAPP chain in a U-shaped conformation with a region of β-strand on either side (in residues 8–17 and 28–37). As the hIAPP monomers stack in along the fibril axis, they form parallel β-sheets as adjacent U-shaped monomers align alongside each other. Two-dimensional infrared spectroscopy (2D-IR) experiments support this proposed structure,[117] as do electron paramagnetic resonance (EPR) experiments.[4] Additional experimental and computational studies have investigated the behavior of amylin and its mutants in the presence of various inhibitors and metals, [6, 122, 83, 121] as well as interactions with other amyloid-forming proteins or with membranes.[48, 47, 75, 123, 122, 29, 21, 66] Furthermore, studies of amylin have extended to examining structural rearrangements during aggregation, [104] as well as identifying aggregation mechanisms.[38]

While mature amylin fibrils have been found to be biologically inert, previous studies have found early-stage aggregates to be associated with cytotoxicity.[103, 50, 18, 82] Prefibrillar species such as dimers, trimers, or higher order oligomers have been proposed as the key species responsible for inducing damage to cell membranes and eventually triggering cell death. [50, 69, 117, 78, 91] Additionally, experiments in which cells undergo addition of hIAPP reveal membrane leakage prior to the formation of mature fibrils [100] and disruption of cell membranes in regions separate from areas of fibril growth,[13] further supporting the link between prefibrillar species and cellular damage.

A thorough investigation of early-stage fibril formation, including the mechanisms underlying amylin aggregation, is therefore critical in building a better understanding of how hIAPP behaves during the onset of disease and whether intermediate structures involved in the aggregation process could potentially be targeted therapeutically. hIAPP residues 20–29 have specifically been proposed to be key in amyloid formation,[120] and mutations to this sequence lead to suppressed amylin aggregation.[120, 79, 92, 1] While residues 20–29 are not located in the parallel β-sheet regions formed in the mature hIAPP fibril, transient parallel β-sheets have been shown via 2D-IR experiments and molecular dynamics (MD) simulations to form in this region prior to fibril formation, particularly in residues 23–27 (with sequence `FGAIL`).[70, 14] Additionally, 2D-IR experiments with dihedral indexing have identified residues L12A13 to form a transient stacked turn or disordered β-sheet during aggregation.[73]

Multiple studies have utilized MD simulations to probe the dimerization of two hIAPP monomers into a U-shaped dimer,[14, 20, 66] including our recent work[38] employing the string method to discover the dimerization mechanism and confirming the formation of transient β-sheet in residues 12–13 and 20–29. The string method-based study found the final U-shaped dimer to be less thermodynamically stable than the disordered dimer by approximately 4.5 k_BT, with a single major free energy barrier of 7 k_BT in the dimerization process associated with formation of an intermediate structure exhibiting transient β-sheet in the aforementioned residues 12–13 and 20–29.[38]

Although these new insights have clarified the dimerization process, many key questions remain with regards to how higher order aggregates are formed, as well as how the unfavorable dimerization of amylin monomers fits into the amylin aggregation process at large. These questions include whether further addition of hIAPP monomers to the growing fibril proceeds similarly to the dimerization mechanism, and whether that process continues to be uphill in free energy. The recent application of the string method to the study of early-stage amylin aggregation has paved the way for studying these higher order aggregates; previous MD-based work on early-stage aggregation have largely been limited to study of the dimer, due to the reliance on the more computationally costly metadynamics approach. In this work, we tackle the next frontier in the early-stage aggregation process—trimerization—by extending the string method approach to discover multiple possible trimerization mechanisms and compare their respective thermodynamic properties. We focus specifically on the comparison of trimerization from three disordered amylin chains versus trimerization from a disordered chain approaching an amylin dimer, the distinctly different free energy barriers encountered in each case, and most interestingly, the role of molecular interactions with water that underlie the key differences between the two aggregation processes.

3.2 Results and Discussion

The finite temperature string method was used, as detailed in the Methods section, to identify and investigate aggregation mechanisms for hIAPP trimerization. We perform the finite temperature string method using two collective variables: (1) a measure of parallel β-sheet character, $\beta^{\text{parallel}}_{\text{RMSD}}$, which is described in the Methods, and (2) the radius of gyration (R_g) of the three protein chains. $\beta^{\text{parallel}}_{\text{RMSD}}$ provides a continuous measure of the amount of parallel β-sheet formed, while R_g provides a measure of compactness; we use them together to characterize the trimerization process from a disordered state with little β-sheet to a more compact aggregated state exhibiting high β-sheet secondary structure.

3.2.1 Trimer Assembly from the Disordered State ("3-chain Assembly")

We begin by investigating assembly of the hIAPP trimer from three separate disordered hIAPP chains; we refer to this assembly process as "3-chain Assembly". Figure 3.1 shows the trimerization pathway discovered via the finite temperature string method. One end of

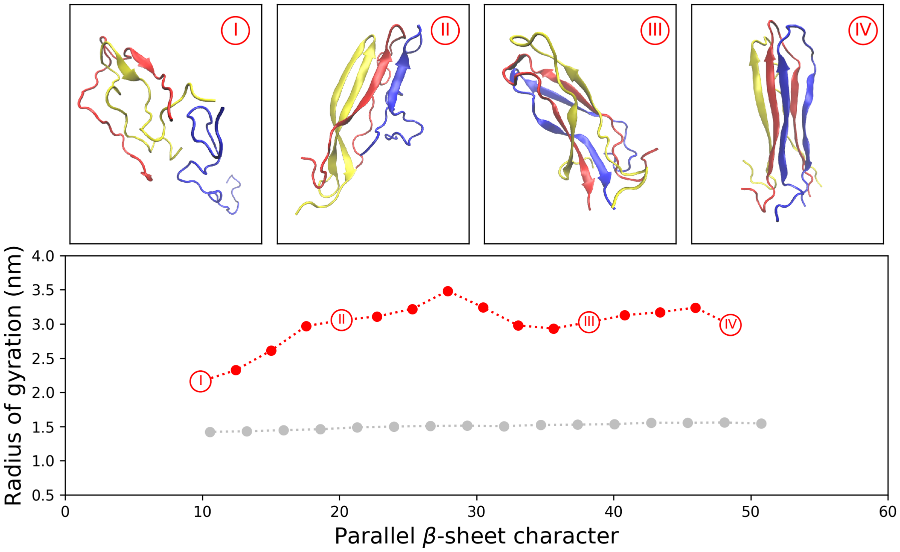


Figure 3.1: 3-chain trimerization pathway calculated from the finite temperature string method, with initial configurations input into the string method shown in grey. Four representative snapshots show mechanistic details during trimer formation. Water and counterions are removed for clarity. The disordered end of the string (Panel I) includes the previously studied dimer intermediate (formed by the yellow and red chains). A similar intermediate comprised of all three chains is discovered and shown in Panel II.

the pathway corresponds to the disordered state of the hIAPP trimer ($R_g = 2.16$, $\beta^{\text{parallel}}_{\text{RMSD}} = 9.87$, shown in Panel I), while the opposite end corresponds to the fully formed trimer ($R_g = 2.99$, $\beta^{\text{parallel}}_{\text{RMSD}} = 48.53$, shown in Panel IV). Representative snapshots highlighting key conformational changes along the trimerization pathway are shown at the top of Figure 3.1.

In Panel I, a disordered amylin chain exhibiting no secondary structure approaches two amylin chains that are loosely associated. Interestingly, the structure of these two chains corresponds to the intermediate structure previously observed for hIAPP dimerization,[38] with parallel β-sheet formed in the turn region suspected to play a key role in amyloid aggregation.[120] Panel II shows that the disordered amylin chain has associated with the dimer intermediate observed in Panel I, forming a similar trimer intermediate that also exhibits parallel β-sheet, this time in the bend region across all three chains. Taken together, Panels I and II suggest that formation of higher-order hIAPP aggregates in the 3-chain assembly scenario proceeds in a stepwise manner, with individual amylin chains added sequentially to a growing aggregate at the intermediate parallel β-sheet stage. Increasing amounts of parallel β-sheet are formed between the three amylin chains through Panels III and IV, leading to the formation of the full trimer, shown in Panel V.

Free energy changes along this 3-chain assembly pathway were calculated as described in the Methods section and are shown in Figure 3.2. A single major free energy barrier of approximately 12 k_BT is observed, corresponding to the formation of the intermediate β-sheet structure as indicated by the snapshots in Figure 3.2 and in Panel II of Figure 3.1. The amylin dimer intermediate rearranges to allow incorporation of the approaching disordered third amylin chain during this 12 k_BT increase in free energy, followed by a slight drop in free energy after the trimer intermediate is formed.

Figure 3.3 shows the changes in average secondary structure per residue over the 3-chain assembly process, calculated using the DSSP algorithm.[53] Through comparison with Figure 3.2, it becomes clear that β-sheet is formed in the intermediate structures in residues 12–13 and 20–29, which have been previously proposed[120, 73] as regions exhibiting transient β-sheet during aggregation and observed to do so in studies of the hIAPP dimer.[38] Figure 3.3 also shows β-sheet forming primarily in the C-termini before advancing to the N-termini, a pattern which was also observed in previous dimer studies.[38]

3.2.2 Trimer Assembly from the Dimer State ("2+1 Assembly")

In addition to studying assembly of the hIAPP trimer from three disordered chains, we applied the finite temperature string method to the study of a trimer assembled from a single disordered amylin chain approaching a fully formed amylin dimer. We refer to this process as "2+1 Assembly". The trimerization pathway discovered using the string method is shown in Figure 3.4; the disordered end of the pathway is found at ($R_g = 2.62$, $\beta^{\text{parallel}}_{\text{RMSD}} = 22.14$, shown in Panel I), and the fully formed trimer is found at the opposite end of the discovered pathway at ($R_g = 2.98$, $\beta^{\text{parallel}}_{\text{RMSD}} = 49.82$, shown in Panel IV). Again, representative snap-

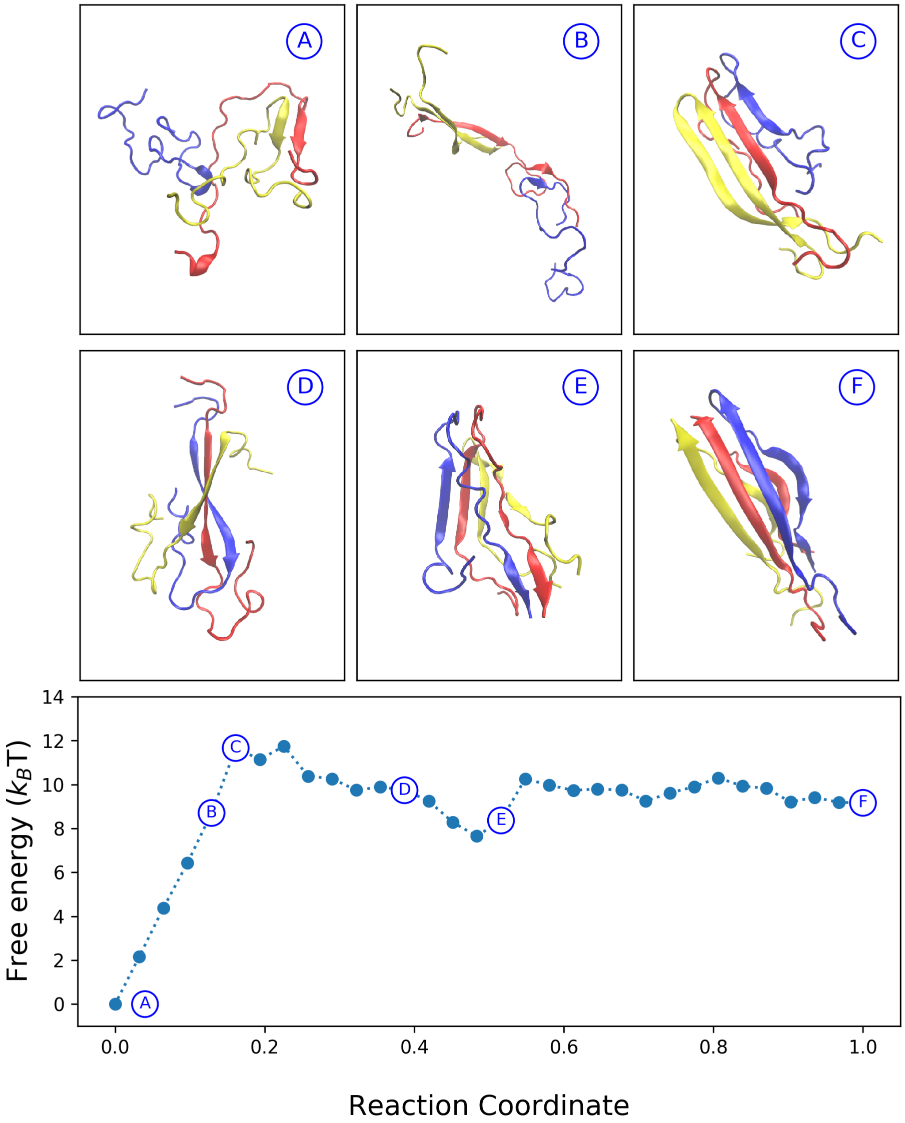


Figure 3.2: Free energy profile along the 3-chain trimerization pathway found via the finite temperature string method. The reaction coordinate extends from 0.0 (disordered state) to 1.0 (fully formed trimer). Free energy is calculated using the procedure described in the Methods, sampling each of the 32 cells for 150 ns each. Representative snapshots are shown to illustrate conformational changes during trimerization. A free energy barrier of approximately 12 k_BT is found, corresponding to the formation of a transient β-sheet structure (shown in Panel C).

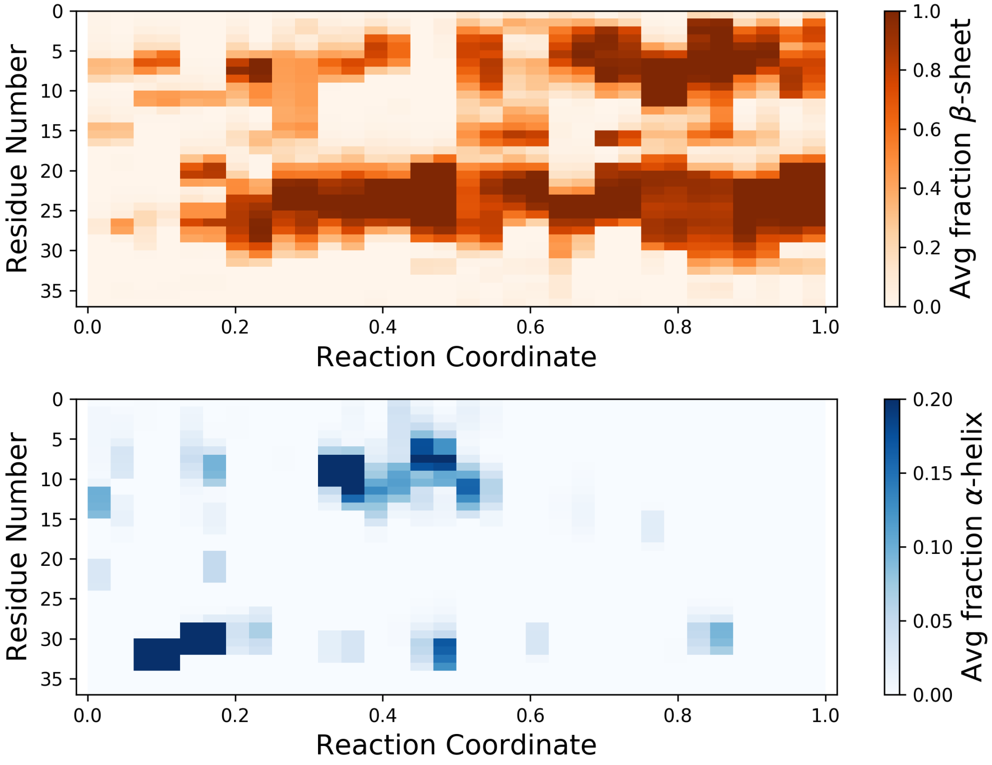


Figure 3.3: Average α-helix and β-sheet secondary structure along the 3-chain trimerization pathway, plotted by residue (ranging from 1 to 37, averaged over all three hIAPP chains). Secondary structure was calculated using the DSSP algorithm[53] and averaged over 150 ns of sampling for each of the 32 bins along the reaction coordinate. β-sheet is found to transiently form in residues 12–13 and 20–29 during formation of the intermediate β-sheet structure, and as trimerization progresses, β-sheet extends to both termini, with more heavily β-sheet character observed in the C-termini.

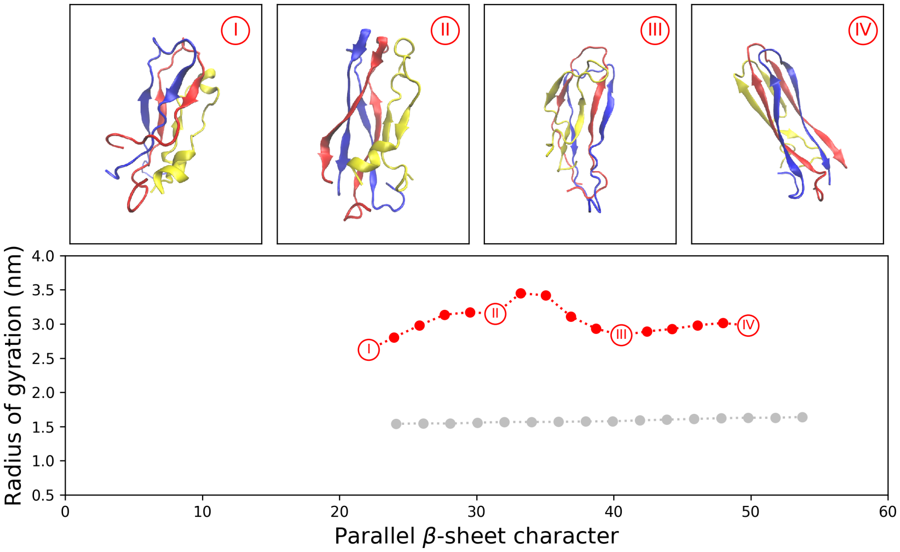


Figure 3.4: 2+1 assembly pathway calculated from the finite temperature string method. Grey points show initial configurations input into the string method. Four representative snapshots illustrate conformational changes observed during trimer formation. Water and counterions are not shown. The disordered end of the string (Panel I) shows a loosely formed dimer and a disordered third chain. Increased β-sheet is formed gradually until the full trimer is formed (Panel IV).

shots highlight the relevant conformational changes along the trimerization pathway, shown at the top of Figure 3.4. Panels I and II show that the hIAPP dimer is first stabilized, before the disordered third amylin chain is gradually incorporated in Panels III and IV. In contrast with 3-chain assembly and previous studies of the dimer,[38] there is no clear intermediate structure exhibiting β-sheet in the bend region.

We also calculate the free energy profile along the 2+1 transition pathway, shown in Figure 3.5. A steady increase of approximately 40 k_BT is calculated for the 2+1 assembly process, with no clear intermediate metastable state. This 40 k_BT rise in free energy is associated with the stabilization of the hIAPP dimer and initial addition of the third hIAPP chain to the dimer, as shown in Panels A-D. The free energy fluctuates around a steady value as the third chain gradually forms greater amounts of β-sheet with the dimer, eventually forming the full hIAPP trimer, shown in Panels E and F.

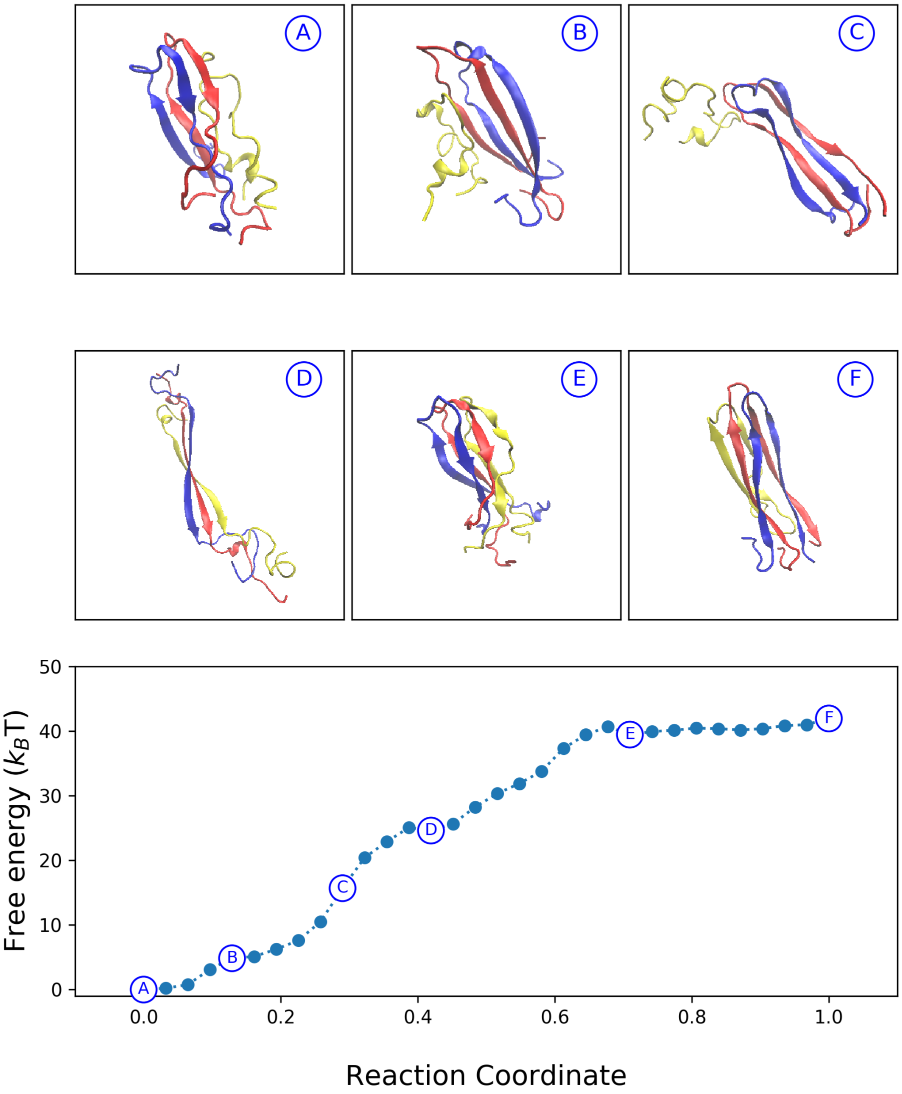


Figure 3.5: Free energy profile along the 2+1 assembly pathway found via the finite temperature string method. The reaction coordinate proceeds from 0.0 (disordered state containing a loosely formed dimer and third disordered chain) to 1.0 (fully formed trimer). Free energy is calculated using the procedure described in the Methods, sampling each of the 32 cells for 150 ns each. Representative snapshots are shown to illustrate conformational changes during trimerization. A free energy barrier of approximately 40 k_BT is found. The intermediate transient β-sheet structure observed during 3-chain assembly is not found.

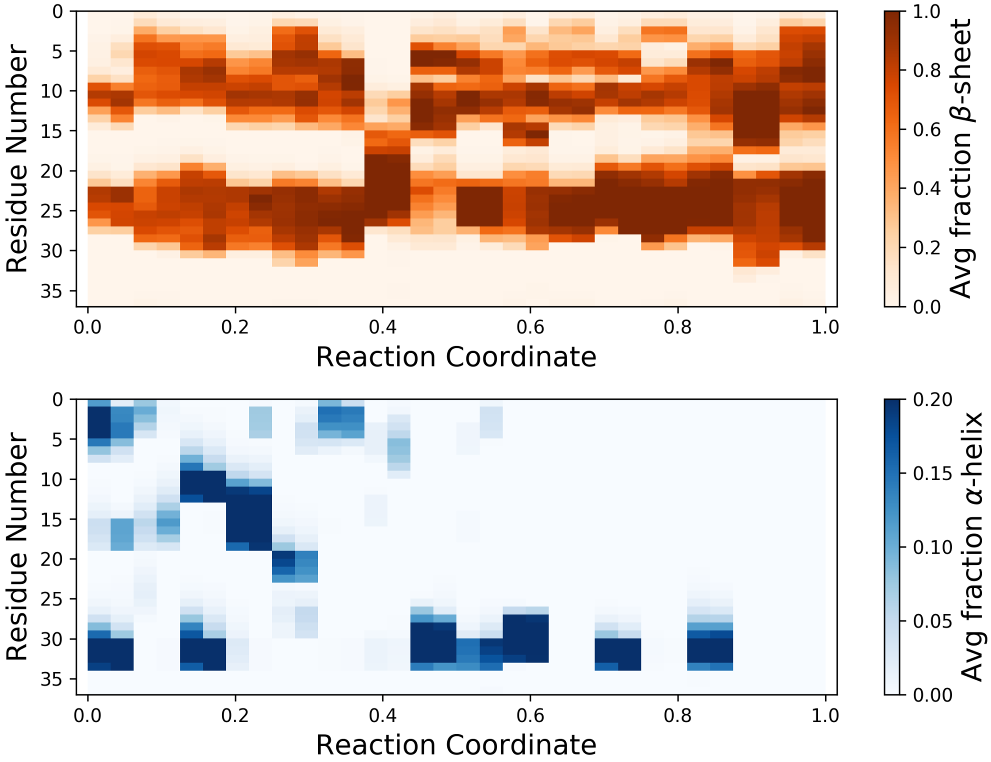


Figure 3.6: Average α-helix and β-sheet secondary structure along the 2+1 assembly pathway, plotted by residue (ranging from 1 to 37, averaged over all three hIAPP chains). Secondary structure was calculated using the DSSP algorithm[53] and averaged over 150 ns of sampling for each of the 32 bins along the reaction coordinate. As with 3+1 assembly, β-sheet extends to both termini, with more heavily β-sheet character observed in the C-termini.

Following our analysis for 3-chain assembly, we now calculate the average secondary structure per residue for the 2+1 assembly process with the DSSP algorithm.[53] Figure 3.6 shows these results, with unsurprising outcomes: compared to 3-chain assembly, 2+1 assembly begins with greater amounts both β-sheet (from the already-assembled dimer) and α-helix (from the approaching third chain). β-sheet gradually increases across the course of 2+1 assembly, with β-sheet forming earlier and more heavily in the C-termini before the N-termini, which was also observed for 3-chain assembly.

3.2.3 "3-chain Assembly" versus "2+1 Assembly"

In order to understand the differences in mechanistic details and the discrepancy between the thermodynamics of the two assembly processes (12 k_BT free energy barrier for 3-chain assembly vs 40 k_BT for 2+1 assembly), we perform a series of comparisons to uncover the key differences between the two trimerization pathways: (1) calculation of protein-water and protein-protein hydrogen bonds during trimerization; (2) decomposition of free energy into entropic and enthalpic components; and (3) decomposition of potential energy contributions into inter- and intra-chain components. These three comparisons will allow us to isolate and contrast specific interactions that contribute to the previously calculated free energy profiles, thereby identifying how the two mechanisms differ and what contributes the unfavorability of 2+1 assembly compared to 3-chain assembly.

We begin by comparing protein-water and protein-protein H-bonds formed during the trimerization process, shown in Figure 3.7. Both assembly processes show a steady rise in protein-protein hydrogen bonds as the full trimer is formed. As expected, 2+1 assembly begins with higher protein-protein H-bond count than 3-chain assembly, reflecting the heavily β-sheet dimer required for the 2+1 pathway. Protein-water H-bonds fluctuate during both trimerization processes; however, there are three distinct downward trends in protein-water H-bonds observed during 3-chain assembly, which is not observed in 2+1 assembly. This is especially distinct during the formation of the full trimer in the last fifth of the 3-chain assembly process, indicating the loss of protein-water H-bonds while protein-protein H-bonds are gained during formation of the full trimer.

The protein-water and protein-protein H-bond profiles shown in Figure 3.7 were then compared with profiles of enthalpic and entropic changes along the two trimerization pathways, shown in Figure 3.8. Free energy profiles for both assembly processes were split into entropic and enthalpic contributions. Entropic contributions were calculated using $\Delta A = \Delta U - \Delta TS$, using potential energy differences ΔU and free energy differences ΔA for each cell sampled along each trimerization pathway, calculated with respect to the first disordered state bin.

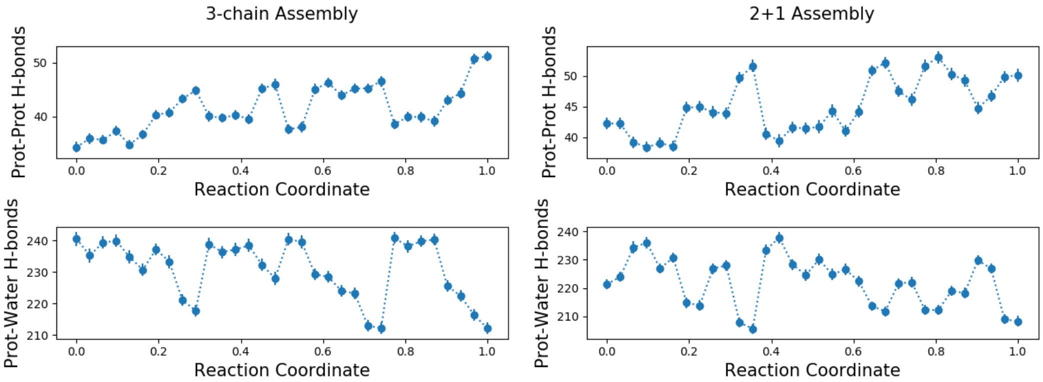


Figure 3.7: Average protein-protein and protein-water hydrogen bonds within 3 Åover the course of dimerization for both trimerization processes, shown with standard error. Average number of H-bonds are calculated from each of the 32 cells sampled during free energy calculation, from 15003 snapshots per cell, using the GROMACS `g_hbond` tool. While protein-protein H-bonds trend upward during both assembly processes, the protein-water H-bonds show a steeper decrease in 3-chain assembly compared to 2+1 assembly. Protein-protein H-bonds are found to form while protein-water H-bonds are lost during 3-chain assembly; this is not observed for 2+1 assembly.

Enthalpic and entropic contributions are both observed to fluctuate throughout 3-chain assembly, with minima in the enthalpic term corresponding with peaks in entropy across the trimerization process. While the free energy profile for 3-chain assembly does not indicate that the fully-formed trimer is metastable, a final dip in the enthalpic term along with its corresponding peak in the entropic term at the end of trimerization suggest there is some degree of entropic stabilization, which occurs in the same period in which protein-protein H-bonds are found to increase at the expense of protein-water H-bonds, shown in Figure 3.7.

In contrast, 2+1 assembly begins with an initial rise in potential energy and decrease in entropy. This feature is not observed in the free energy decomposition for 3-chain assembly, suggesting that the 2+1 assembly process must undergo a transition through enthalpically unfavorable conditions that are not observed during 3-chain assembly. As trimerization progresses after the third disordered chain meets the already-formed dimer structure, the entropic contribution steadily rises, leading to a steady and substantial rise in free energy of approximately 40 k_BT, which was previously shown in Figure 3.5.

Additional insights into the differences between 3-chain assembly and 2+1 assembly were

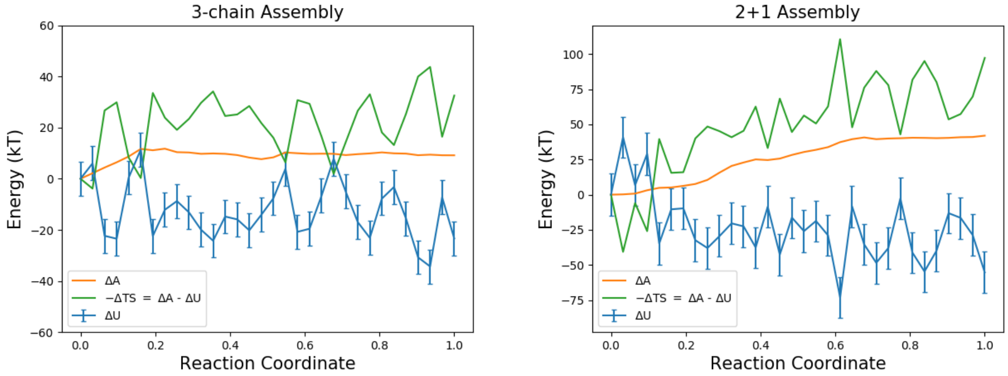


Figure 3.8: Decomposition of free energy into enthalpic (ΔU) and entropic ($-\Delta TS$) contributions for both trimerization mechanisms. ΔA is taken from the free energy calculation, performed as described in Methods and shown in Figures 3.2 and 3.5. Average potential energy is calculated from each of the 32 cells sampled during free energy calculation, from a total of 15003 snapshots per cell. Entropic contributions are then calculated as $-\Delta TS = \Delta A - \Delta U$.

found by further decomposing the potential energy profiles for the two assembly processes into intrachain, interchain, and chain-water contributions. The stark differences in trends between the two assembly processes are shown in Figure 3.9. In 3-chain assembly, intrachain potential energy steadily rises throughout trimerization, while in 2+1 assembly it steadily decreases. Meanwhile, interchain potential energy fluctuates during 3-chain assembly before finally stabilizing during formation of the full trimer, in the same period of time where protein-water H-bonds sharply decreased while protein-protein H-bonds sharply increased (Figure 3.7). In contrast, interchain potential energy is initially stable for 2+1 assembly, rises during trimerization, and fluctuates during formation of the full trimer. Chain-water potential energy appears to fluctuate independently of inter- and intrachain interactions during 3-chain assembly, with a slight upward trend as the trimer forms; in 2+1 assembly, this chain-water potential energy exhibits a slight downward trend, with dips in chain-water interactions corresponding to peaks in intrachain interactions during the first half of trimerization, and with interchain interactions during the last half.

Taken together, the comparisons above indicate that the role of water in each trimerization process contributes greatly to the distinct differences between 3-chain and 2+1 assembly;

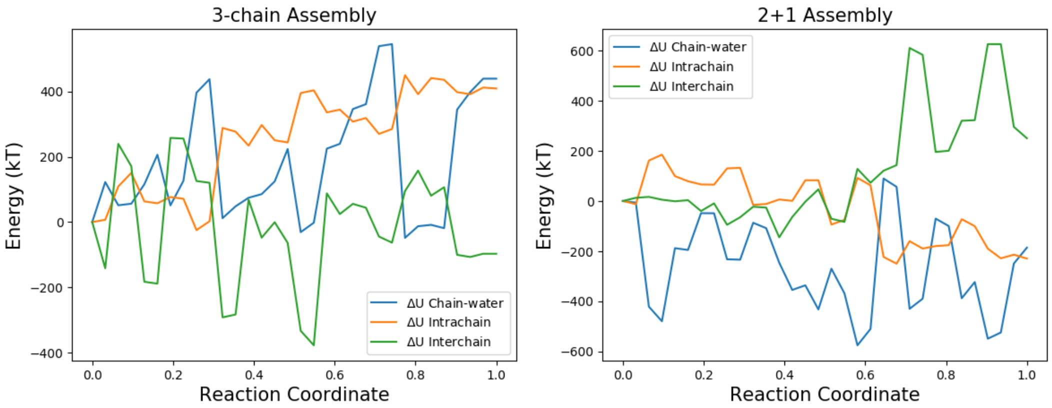


Figure 3.9: Decomposition of potential energy into interchain, intrachain, and chain-water interactions for both trimerization mechanisms. Energies are calculated from each of the 32 cells sampled during free energy calculation, from a total of 15003 snapshots per cell. Note the contrasting trends between the two assembly processes for all three interactions plotted.

specifically, the pre-formed dimer in 2+1 assembly is stabilized by its protein-water interactions, and breaking of these protein-water interactions in order to form new protein-protein contacts during trimerization becomes unfavorable compared to formation of protein-protein contacts from three disordered chains. Decreases in protein-water H-bonds tend to correspond to gains in protein-protein H-bonds in 3-chain assembly, but this is not observed during 2+1 assembly. Increased enthalpic and decreased entropic contributions to free energy are observed in the first portion of 2+1 assembly, when the pre-formed dimer and disordered third chain are still separated and each surrounded by water; these conformations are not observed at any time during the 3-chain assembly process, and neither are these thermodynamic features. A sharp rise in entropic contributions and fall in enthalpic contributions, however, is observed at the very end of 3-chain assembly, corresponding to the final simultaneous drop in protein-water H-bonds and gain in protein-protein H-bonds; this, in turn, is not observed at any time during the 2+1 assembly process. Furthermore, chain-water potential energy decreases during 2+1 assembly, with minima corresponding to peaks in either intra- or interchain potential energy, indicating a tendency toward stabilization of protein-water interactions during 2+1 assembly at the expense of establishing energetic stabilization of protein-protein interactions within the forming amylin trimer. As the pre-formed dimer and

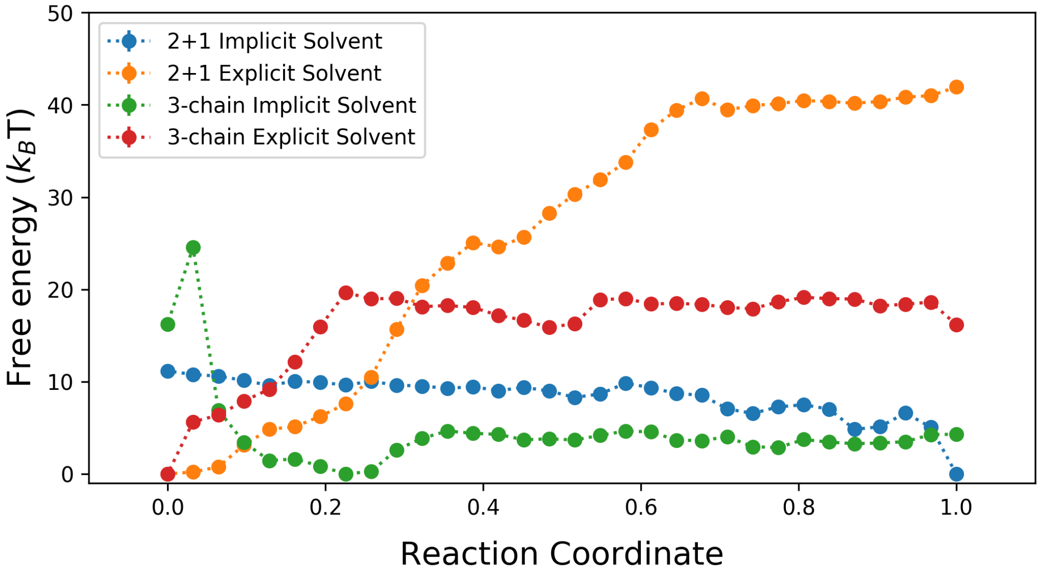


Figure 3.10: Comparison of free energy profiles for both 3-chain and 2+1 assembly, using an explicit water model (as was shown in Figures 3.2 and3.5) versus implicit water. When the explicit water interactions are replaced with a dielectric continuum as in the implicit model, the free energy profiles are noticeably flattened, suggesting that the moleuclar-level interactions with water are responsible for the free energy barriers originally observed.

disordered third chain in 2+1 assembly are brought together in water, the stabilization of pre-existing protein-water contacts is prioritized before the establishment and stabilization of new interactions between all three chains, reflected in the thermodynamic quantities discussed and compared here. We further confirm that it is the molecular interactions involving water that drive the differences observed in 3-chain and 2+1 assembly by recalculating the free energy profile of trimerization for both processes (as described in the Methods) with explicit water molecules replaced by the OBC GBSA implicit water model.[85] With every individual interacting water atom now replaced with a dielectric continuum, the free energy profiles are considerably flattened, as displayed in Figure 3.10, suggesting that the key differences between 3-chain and 2+1 assembly indeed originate from molecular-level interactions with water.

3.3 Conclusions

A finite-temperature string method approach was used to study multiple pathways of hIAPP trimer formation and their thermodynamics, with specific focus on assembly from three disordered chains ("3-chain assembly") versus assembly from a dimer and one disordered chain ("2+1 assembly"). In both 3-chain assembly and 2+1 assembly, the fully-formed trimer was found to lie in a global free energy minimum, separated from the fully-formed dimer by a climb in free energy. This climb is approximately 12 k_BT for 3-chain assembly, and a steep 40 k_BT for 2+1 assembly; neither fully-formed trimer structure is found to be metastable.

For 3-chain assembly, crossing of the 12 k_BT barrier corresponds to formation of an intermediate structure exhibiting parallel β-sheet in residues 12–13 and 20–29 across all three hIAPP chains, a structure which has been previously proposed and similar to the intermediate β-sheet structure observed in computational studies of the hIAPP dimer. Interestingly, the string method identifies the disordered end of the 3-chain assembly as a conformation which includes this dimer intermediate state, suggesting that hIAPP aggregation via 3-chain assembly proceeds in a stepwise manner, using the intermediate β-sheet structure as a template for fibril propagation.

Furthermore, a series of comparisons was used to investigate the nature of the stark difference between the 12 k_BT 3-chain assembly process and the 40 k_BT 2+1 assembly process. Analysis of protein-water and protein-protein H-bonds over trimerization, decomposition of free energy into enthalpic and entropic contributions, and decomposition of potential energy into interchain, intrachain, and chain-water interactions linked the key differences between 3-chain and 2+1 assembly to their differences in molecular-level interactions with water. The relatively high number of pre-existing chain-water interactions in 2+1 assembly compared to 3-chain assembly underlie the individual differences in entropic, enthalpic, and hydrogen bond contributions, which together ultimately result in the two distinctly different trimerization free energy profiles.

Although the current work demonstrates that 3-chain assembly is more thermodynamically favorable versus 2+1 assembly, both processes are uphill in free energy, along with the dimerization process studied in our previous work. However, aggregates have been demonstrated to form experimentally through seeding and incubation procedures. Based on our finding that systems with an increased number of protein-water H-bonds undergo less favorable aggregation, we hypothesize that aggregates can be stabilized as a result of increased competition for hydrogen bonding with water from other molecules or due to a densely concentrated environment. A forthcoming publication will investigate this further by introducing additional species into the system, including salts and readily H-bond-forming molecules.

Additionally, questions still remain on whether further growth toward higher order oligomers will remain uphill in free energy, and whether spontaneous fibril growth only takes place once a certain-sized hIAPP oligomer is formed. As we look toward higher order aggregates, questions arise about the formation of larger oligomers and whether these processes will proceed in a similar manner as dimer and 3-chain trimer formation, or perhaps a more complex process due to the increased number of monomers involved. Larger systems are being studied to further clarify these issues, and will be a target of future work.

3.4 Methods

3.4.1 Human Amylin Trimer

The hIAPP trimer system was designed based on the system used previously for hIAPP dimer simulations test by [38]. The amino acid sequence for hIAPP is KCNTATCATQR-LANFLVHSSNNFGAILSSTNVGSNTY. Each C-termini is amidated, and Cys2 and Cys7 on each chain is linked by a disulfide bond. Protonation states were assigned on the basis of pKa values in water at a pH of 7.0; each hIAPP chain has a formal charge of +3, and chloride counterions are included to ensure charge neutrality. We use the AMBER ff99SB*-

ILDN force field,[65, 64, 88] which was chosen for its previously demonstrated ability to accurately capture behavior of amyloidogenic polypeptides and other intrinsically disordered proteins.[46, 32] The protein system was placed in a periodic cubic box with side length 15.0 nm with 110,008 TIP3P water molecules.[52] Volume was kept constant, with coulombic forces calculated via the particle mesh Ewald algorithm [25, 31] and temperature held at 298 K using the Nosé-Hoover thermostat.[84] A timestep of 2 fs was used, and hydrogen bond lengths were constrained to equilibrium values using the LINCS algorithm.[43]

3.4.2 Finite Temperature String Method

We study hIAPP trimerization by employing the finite-temperature string method [115], which calculates a transition pathway using a set of local points ("nodes") connected in series by a smooth curve ("string") in collective variable space. For the trimer system, we use two intuitive collective variables: (1) parallel β-sheet character $\beta^{\text{parallel}}_{\text{RMSD}}$, defined below in Equation 3.1; and (2) the radius of gyration R_g of the three hIAPP chains, which provides a measure of spatial distance between each hIAPP monomer.

The parallel β-sheet character of a particular amino acid sequence between residues indices u and v is defined as:[90]

$$\beta^{\text{parallel}}_{\text{RMSD u-v}} = \sum_{\beta} \frac{1 - \left(\frac{\text{RMSD}}{0.8\text{Å}}\right)^{8}}{1 - \left(\frac{\text{RMSD}}{0.8\text{Å}}\right)^{12}} \tag{3.1}$$

Equation 3.1 sums over every possible pair of three-residue segments bounded by residues u and v in each hIAPP monomer. "RMSD" refers to the root mean square deviation (in Å) of the positions of the N, C_α, C_β, C, and O backbone atoms of the residues in each pair of three-residue segments from those in an ideal parallel β-sheet. $\beta^{\text{parallel}}_{\text{RMSD u-v}}$ essentially measures the number of three-residue pairs that are arranged similarly to the configuration of an ideal parallel β-sheet.

Each string is discretized into 16 nodes, with each node's location in collective variable space denoted by $\mathbf{z}_\alpha$, where α is the node index along the string ($\alpha = 0, 1, ..., 15$). The string nodes are used to generate a Voronoi tessellation, where each node is associated with a corresponding Voronoi cell, which consists of the region in CV space closer to its associated string node than any other node along the string. We assume Euclidian geometry for this collective variable space. At every iteration of the string method, each Voronoi cell is sampled such that there is no bias applied while the system is within the boundaries of the Voronoi cell; however, if the system departs from the Voronoi cell, a soft wall harmonic restraining potential is applied:

$$V_{\text{Voronoi}} = \begin{cases} 0 & \text{system in cell} \\ k_i(\|\mathbf{z}(\mathbf{x}(t)) - \mathbf{z}_\beta\|)^4 & \text{system out of cell, in cell } \beta \end{cases} \tag{3.2}$$

Each Voronoi cell is sampled for 100 ps per string method iteration, and the running average of each node's explored location in collective variable space $\overline{\mathbf{z}_\alpha}$ is tracked starting from the first string method iteration. The string is updated every nth iteration according to:

$$\mathbf{z}_\alpha^{n+1} = \mathbf{z}_\alpha^n - \Delta\tau(\mathbf{z}_\alpha^n - \overline{\mathbf{z}_\alpha}) + \mathbf{r}_\alpha \tag{3.3}$$

where we choose $\Delta\tau$ to be 0.1, smoothing parameter $\mathbf{r}_\alpha$ to be 0 for nodes on each end of the string ($\alpha = 0$ or 15), and for the interior nodes:

$$\mathbf{r}_\alpha = \kappa N^2 \Delta\tau(\mathbf{z}_{\alpha+1} + \mathbf{z}_{\alpha-1} - 2\mathbf{z}_\alpha) \tag{3.4}$$

with smoothing parameter κ chosen to be 0.1 and the total number of nodes on the string N is 16. Following every string update, a cubic spline interpolation is drawn through the 16 nodes, and the nodes are then redistributed along the string in order to maintain equal arc-lengths between adjacent nodes. These steps are iterated until the string converges to a final pathway.

Upon convergence, the free energy is computed along the final string by calculating π_α, the equilibrium probability of the system to be found in Voronoi cell α, which is then used to calculate the corresponding free energy A_α [74]:

$$A_\alpha = \frac{1}{k_B T} \log(\pi_\alpha) \tag{3.5}$$

To improve resolution of the resulting free energy profile, we further discretize the original 16 node string to a total of $N = 32$ Voronoi cells. Each of the Voronoi cells is sampled using the same soft wall restraints described in Equation 3.2, for 50 ns for multiple runs. For each system sampling cell α, we collect T_α, the total simulation time spent within cell α, as well as $N_{\alpha\gamma}$, the number of times the system escapes into a neighboring cell γ. The equilibrium probabilities π_α are calculated with the following system of equations, where $\nu_{\alpha\gamma} = \dfrac{N_{\alpha\gamma}}{T_\alpha}$ is the rate of escape from cell α into γ:

$$\sum_{\gamma=1}^{N} \pi_\gamma \nu_{\gamma\alpha} = \sum_{\gamma=1}^{N} \pi_\alpha \nu_{\alpha\gamma} \tag{3.6}$$

$$\sum_{\alpha=1}^{N} \pi_\alpha = 1 \tag{3.7}$$

String method simulations were performed using the GROMACS 4.6.7 simulation package [5, 44], the PLUMED 2.1 plugin [12], along with custom code to perform string method calculations.

CHAPTER 4

DYNAMICS OF PEPTIDE AMPHIPHILES FOR PHOSPHATE CAPTURE AND RELEASE

The recovery of valuable resources from wastewater is becoming increasingly important as global population rises and natural resources are depleted. One such resource is phosphate, which is critical for its use in fertilizers in maintaining food production worldwide and lacks any viable substitute. Biologically-inspired peptide amphiphiles are a particular type of material that can address this goal of sequestering phosphate from wastewater, by incorporating a phosphate-binding peptide sequence with an alkyl chain that drives self-assembly to form a self-assembling micellar structure with phosphate-sequestering properties. In this work, we investigate the preliminary peptide amphiphile candidate C16GGGhex, which is made up of a 16-carbon alkyl tail connected to a known pH-responsive phosphate-binding hexapeptide via a 3 glycine linker. We use a combination of molecular dynamics and enhanced sampling methods to study the potential of C16GGGhex for efficient phosphate capture and release at high and low pH conditions. Screening and clustering calculations show that phosphate may bind with C16GGGhex at multiple locations along its peptide region, not solely at the known phosphate-binding hexapeptide. Adaptive biasing force (ABF) simulations of both single C16GGGhex chains and a flat layer of C16GGGhex indicate preferential binding of phosphate at low pH, with three distinct phosphate-binding locations identified in single-chain studies, while no preferential binding is observed at high pH.

4.1 Introduction

Sustainable access to clean water is an urgent global priority and involves wide-ranging challenges, including establishment of access to clean water for billions of people, as well as detection and removal of harmful toxins and contaminants from the water supply. An attractive solution for addressing both water scarcity and water purification in a sustainable manner is the development of advanced water treatment technologies that simultaneously decontaminate wastewater while recovering valuable nutrients and resources for reuse.[76, 35, 37, 24] Phosphate is one particular resource for which this technology would be especially impactful, due to its critical role in sustaining food production worldwide and the dwindling global availability of non-renewable natural phosphate deposits.[24] The lack of viable substitutes for phosphate in agricultural applications have ignited efforts in developing methods for recycling phosphate from wastewater and agricultural runoff, from which excess phosphate currently cannot be recovered and instead contributes to the dangerous overgrowth of algae.[95]

Interest in developing biomimetic approaches to phosphate sequestration has lead to the investigation of various peptide sequences with the ability to bind to phosphate. One particular motif that has been shown to bind phosphates is known as the “P-loop”,[99, 118, 77] characterized by the sequence Gly-Xxx-Xxx-Xxx-Xxx-Gly-Lys-(Ser,Thr), which forms a nest-like conformation and has been shown to bind to phosphate in the body.[99] Furthermore, screening of thousands of phosphate-binding proteins have identified each of the amino acids found in the P-loop motif as frequently occurring amino acids in the screened phosphate-binding sites, in addition to arginine, aspartic acid, and glutamic acid.[36] Building upon study of the P-loop, Bianchi et al. have designed a hexapeptide with the sequence Ser-Gly-Ala-Gly-Lys-Thr (SGAGKT),[8] which was demonstrated to selectively bind to phosphate above pH 6, via a nest-like conformation with hydrogen bonds between the phosphate and the NH backbone atoms.

One way in which this hexapeptide designed for phosphate-sequestration can be incorporated into functional materials is by conjugating the peptide sequence to an alkyl tail to form a peptide amphiphile (PA). PAs self-assemble into micellar assemblies in water, oriented such that the hydrophilic peptide regions face outward, shielding the hydrophobic alkyl regions that aggregate in the center of the micelles.[114] The peptide sequences and alkyl tails can be strategically tuned to impart specific functionalities, which has lead to their successful deployment in various biomedical applications, including drug delivery,[3] immunotherapy,[11] and medical imaging.[15]

Previous work includes the design of the PA C16GSH,[63] a branched PA with a C16 tail attached to two peptide chains, which forms a pH-reversible structure due to the inclusion of histidine residues in one of the branches. At pH above 6.5, C16GSH assembles into a self-supporting hydrogel network of entangled wormlike micelles, while at lower pH, the assembly becomes liquid-like. The pH range in which the hydrogel forms falls within the range in which the SGAGKT sequence was observed to bind to phosphate, making a mixture of the two PAs an attractive design candidate for a pH-controllable structure that can be switched between gel-like and liquid-like states, corresponding to phosphate capture and phosphate release.

The vast design space for such a PA presents a challenge, and we begin by investigating a simple preliminary candidate PA that incorporates the C16 tail from C16GSH with the SGAGKT hexapeptide that has been shown to successfully sequester phosphate; these two components are linked together with a sequence of 3 glycines to form a newly designed PA C16GGGhex (Figure 4.1). A detailed understanding of this preliminary candidate PA is valuable for uncovering the key parameters governing successful phosphate-binding of a peptide sequence once incorporated into a larger PA system and will ultimately provide guidelines for design and optimization for successful recycling of phosphate from wastewater. Here, we focus on computational characterization of C16GGGhex and its phosphate-binding properties; an accompanying paper detailing experimental characterization will be prepared separately.

In this work, we use atomistic molecular dynamics with the adaptive biasing force sampling method to study the phosphate-binding dynamics of C16GGGhex, with focus on differences in binding behavior at high and low pH conditions. We begin with an evaluation of binding dynamics for a phosphate molecule binding to a single C16GGGhex chain at both pH conditions, and compare these results to previous studies of the hexapeptide alone. We then scale the system up to a periodic layer of C16GGGHex to evaluate the thermodynamics of phosphate binding with multiple C16GGGhex organized in an assembled structure,

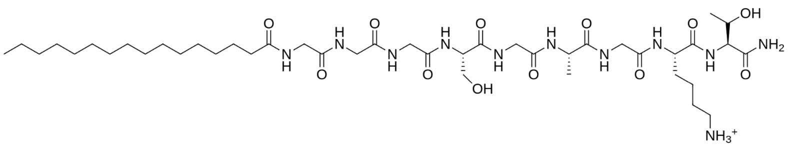


Figure 4.1: Molecular structure of C16GGGhex. A C16 alkyl tail is connected by a 3 glycine linker to the SGAGKT hexapeptide, which has previously been demonstrated to bind to phosphate in a pH-responsive manner.

and compare these results to the single-chain studies. Interestingly, we find that once the hexapeptide is incorporated into the PA structure, the pH dependence of phosphate binding is reversed, with preferential phosphate capture occurring at low pH and no preferential binding observed at high pH conditions. Furthermore, the phosphate ion is found to associate along the peptide region of the PA in a delocalized manner, with three separate free energy minima identified along both the glycine linker and SGAGKT region.

4.2 Results and Discussion

We apply a combination of brute force molecular dynamics (MD) and the adaptive biasing force (ABF) enhanced sampling method to investigate phosphate-binding behavior of C16GGGhex at both high and low pH conditions. Simulation tools and details are described in the Methods section. High and low pH conditions are modeled by appropriately changing the protonation state of the phosphate; we model low pH based on the expected phosphate species at observed at pH 6 ($H_2PO_4^-$) and high pH on the that for pH 11 (HPO_4^{2-}). The peptide end of the C16GGGhex PA is amidated; its structure is shown in Figure 4.1.

4.2.1 Phosphate Binding to a Single C16GGGhex Chain

We begin by screening both pH conditions using brute force MD simulations, in order to gain basic understanding of the configurations assumed by the bound C16GGGhex-phosphate structure. This information is then used to choose effective collective variables (CVs) to describe the system dynamics, which are necessary in the ABF enhanced sampling simulations used to study the thermodynamics of binding at each pH condition. Brute force MD simulations were initialized from 9 different starting configurations for each pH condition, for a total of 180 ns simulated at each pH. Snapshots were collected every 5 ps for a total of 36001 snapshots for each system, which were each then rotationally and translationally aligned by the protein coordinates in each snapshot. The aligned snapshots were then analyzed using the GROMACS `cluster` tool, using the gromos clustering algorithm with a cutoff of 0.27 nm,[27] producing a total of 177 clusters for low pH and 2517 clusters for high pH. The discrepancy in the number of clusters discovered at each pH is linked to the greater number of unbound snapshots at high pH, which is our first indication that the propensity for phosphate binding is distinctly different at the two pH conditions.

Before using the ABF sampling method to investigate this potential difference in binding likelihood, we must choose appropriate CVs that characterize the dynamics of phosphate-binding in our system. To do this, we examine the 8 most populated clusters calculated from each pH. These snapshots are displayed in Figure 4.2. As expected, we identify clusters corresponding to binding with the SGAGKT hexapeptide, which can be identified in the snapshots as structures where the phosphate is surrounded by a 3-prong "claw" made up of hexapeptide side chains.

However, by visual examination, we find that that the phosphate associates with the C16GGGhex peptide region in multiple regions, and that the bound phosphate is not solely localized to the SGAGKT hexapeptide region. Clustering results show that the phosphate may also attach at the GGG linker, as well as in between, contacting both GGG and SGAGKT regions. Based on this finding that the bound phosphate is delocalized around the PA's entire peptide region, we choose two distance CVs to characterize single-chain phosphate binding: (1) d_{SGAGKT}, distance from the phosphate to the center of the 3-prong SGAGKT binding pocket; and (2) d_{GGG}, distance from the phosphate to the center of the GGG binding region. Details on how these distances are calculated are found in the Methods

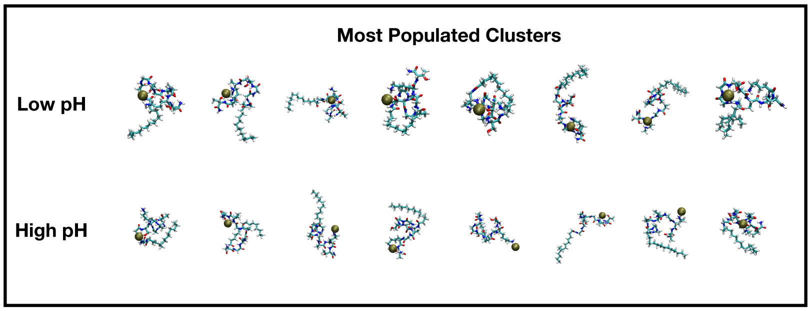


Figure 4.2: Top 8 most occupied clusters obtained for both high and low pH single-chain phosphate binding simulations. Populations of clusters decrease from left to right. Note the multiple locations at which phosphate is found to associate with the C16GGGhex chain; this motivates our choice in collective variables described in the text.

section.

Using these two distance CVs, we then perform ABF simulations to calculate the free energy landscapes of a phosphate ion binding to a single C16GGGhex chain at both high and low pH. The two distinctly different free energy landscapes are shown in Figure 3 and 4. For binding at low pH, three free energy minima are found: (1) at $d_{\mathrm{SGAGKT}} = 0.12$, $d_{\mathrm{GGG}} = 0.71$; (2) at $d_{\mathrm{SGAGKT}} = 0.18$, $d_{\mathrm{GGG}} = 1.19$; and (3) at $d_{\mathrm{SGAGKT}} = 0.57$, $d_{\mathrm{GGG}} = 1.98$.

The free energy differences between the three minima are moderate, with the deepest minima at (1) and the most shallow at (2). Minima (2) lies at a smooth 41.6 kJ/mol climb in free energy above (1), while a 42.7 kJ/mol barrier must be passed going from minima (2) to minima (3), with a net free energy difference of -11.7 kJ/mol. All three minima, however, lie at a much lower free energy when compared to the free energy maximum found in the upper right hand side of the free energy landscape, corresponding to the a unbounded state where the phosphate is located far from both the GGG linker and the SGAGKT hexapeptide; this is found at $d_{\mathrm{SGAGKT}} = 1.07$, $d_{\mathrm{GGG}} = 1.51$, at 99.0 kJ/mol above the highest free energy minima (2). Taken together, this suggests that binding between C16GGGhex and phosphate is favorable at low pH, with three preferential binding regions along the the entire GGGSGAGKT segment, which can be interchanged by crossing moderately high free energy barriers.

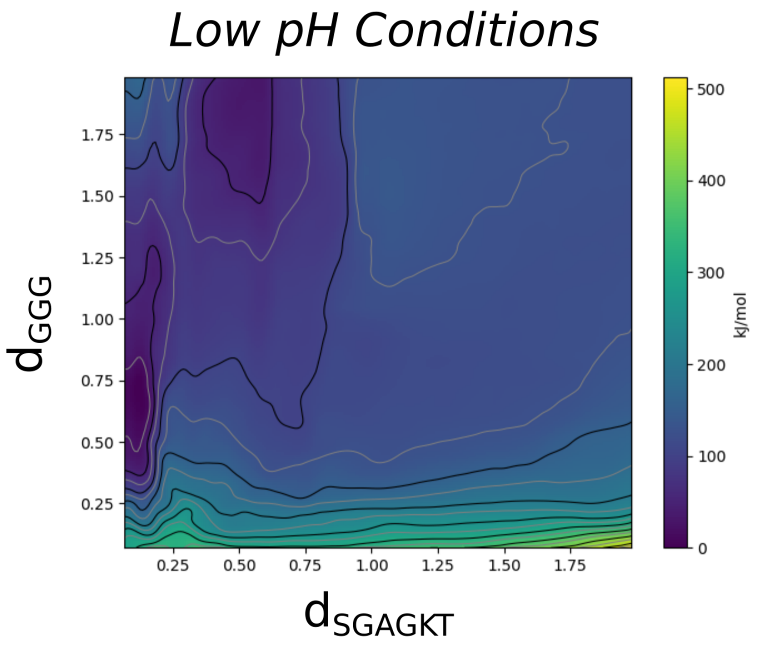


Figure 4.3: Free energy surface for single-chain phosphate binding at low pH conditions. Three distinct free energy minima are found for the low pH system, with moderate free energy barriers between them.

Unlike at low pH, there are no distinct free energy minima associated with preferential phosphate-binding regions at high pH. Instead, a wide free energy well is found to include configurations where the phosphate is both near and far from the C16GGGhex chain, with free energy only decreasing as distance between the phosphate and C16GGGhex chain grows, suggesting C16GGGhex will fail to sequester phosphate at high pH conditions. Interestingly, these results follow an opposite pattern from previous studies of the SGAGKT hexapeptide alone, which was shown to bind phosphate at pH of 6 and above, suggesting that interactions with added GGG linker or the alkyl tail may drive more favorable binding to occur at lower pH conditions. Our observation via simulation that phosphate binding is favorable at low pH and not at high pH is further corroborated by our experimental findings that phosphate is bound at pH 6 and released at pH 11, with switching between binding and release occurring within seconds of pH adjustment; these results are being prepared as part of a forthcoming manuscript.

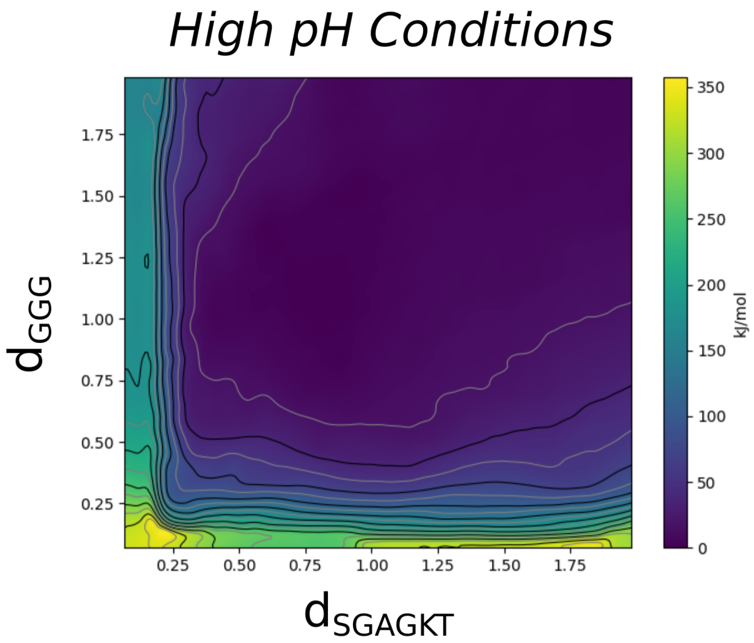


Figure 4.4: Free energy surface for single-chain phosphate binding at high pH conditions. Unlike the low pH free energy landscape in Figure 4.3, there are no clear minima associated with bound phosphate. Free energy steadily decreases as the phosphate moves further away from the C16GGGhex chain.

4.2.2 Phosphate Binding to a Flat Layer of C16GGGhex

Our ultimate goal is to design a PA system that is able to sequester phosphate and release it in a controllable manner, as part of a self-assembled PA structure. Thus, it is necessary to extend the single-chain studies of C16GGGhex and its phosphate-binding properties to a larger system in which phosphate may be captured by an assembled state of C16GGGhex. While we expect C16GGGhex to form long cylindrical micelles, we choose here to first perform the less computationally expensive studies of phosphate binding to a flat periodic layer of C16GGGhex, with the goal of gaining a strong basic understanding of phosphate binding to multiple C16GGGhex chains in a relatively simple geometry before moving on to a system of flexible long micelles.

For both high and low pH conditions, a flat periodic layer of C16GGGhex was prepared, with 25 C16GGGhex chains per 3-dimensional rectangular simulation box. The C16GGGhex chains are initialized in parallel configurations along the z-axis of the simulation box, with

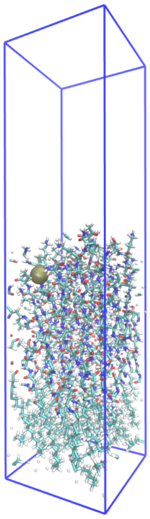


Figure 4.5: Schematic showing setup of flat layer C16GGGhex simulations. 25 C16GGGhex chains are distributed evenly across the x-y plane of the box, with the tail-end alkyl carbons anchored at z = 0. The ABF method is then used to drive insertion of the phosphate to multiple depths within this flat C16GGGhex layer in order to calculate the associated free energy profile.

the z-coordinate of each tail-end alkyl carbon near $z = 0$. The box was then solvated with water and equilibrated with position restraints on the C16GGGhex layer in order to stabilize the layer at the bottom of the box. Further details are described in the Methods section. The 25 C16GGGhex chains are distributed across the 2.5 nm × 2.5 nm x-y area of the box, for an average concentration of 4 chains per nm^2. A snapshot of the system is shown in Figure 4.5, with waters removed for clarity. We perform these simulations at constant volume and temperature in the absence of pressure-coupling, which causes the C16GGGhex chains begin to form cone-like assemblies rather than a layer-like structure, rendering the ABF calculations unhelpful.

We then perform ABF simulations using distance between the z-coordinate of the center of mass of the phosphate ion and the z-coordinate of the center of mass of every tail-end alkyl carbon as a collective variable, effectively measuring the depth at which the phosphate is able to penetrate into the C16GGGhex assembly. The free energy profiles obtained from each pH condition are shown in Figure 4.6.

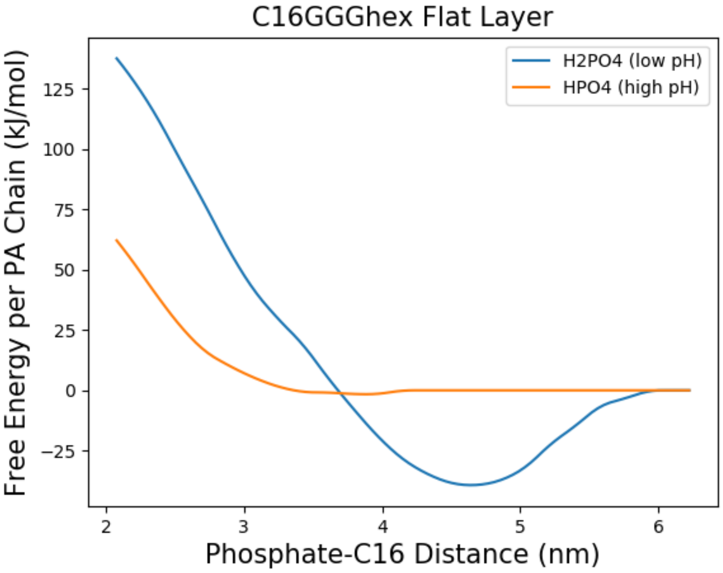


Figure 4.6: Free energy profile for phosphate binding to a flat layer of C16GGGhex. A clear free energy minima is found for low pH conditions, spanning a width of approximately 2 nm; this is in agreement with our single-chain results, which indicated that phosphate is able to bind at low pH in a variety of locations along the GGGhex segment. The high pH free energy profile shows a very small minima of 2.5 kJ/mol, indicating a small amount of stability imparted by the insertion of phosphate into a dense assembly of C16GGGhex; besides this feature, there is no indication that phosphate binding occurs at high pH.

The two free energy per chain profiles are in agreement with our findings for the single-chain phosphate-binding scenario, with binding at low pH exhibiting a clear free energy well while binding at high pH does not. The free energy at the low pH minimum lies 39.7 kJ/mol per chain below that of the unbound state. In comparison, the high pH system exhibits a small minimum 2.5 kJ/mol per chain, which is barely perceptible when compared next to the low pH free energy profile. Single-chain calculations at high pH indicated no free energy minima, suggesting that this mildly stable bound state arises purely from the effects of multiple C16GGGhex chains in close contact. Free energy increases steeply for both pH conditions as the distance between the phosphate and the tail-end aklyl carbons decreases, indicating unsurprisingly that movement of the phosphate beyond the peptide region of the PA assembly is highly unlikely. Futhermore, the free energy minima for low pH binding is broad, spanning across 2 nm in width, which aligns with our observation in single-chain binding that the location of the phosphate is delocalized along the peptide region once bound.

4.3 Conclusions

A combination of brute force molecular dynamics and adaptive biasing force simulations were used to investigate the phosphate-binding properties of the peptide amphiphile C16GGGhex at high and low pH conditions. Single-chain binding to phosphate was first screened with a series of brute force MD simulations, and configurations were then taken from the screening simulations and passed through a clustering algorithm to identify characteristic C16GGGhex-phosphate configurations. Results from clustering indicated that phosphate is able to bind to C16GGGhex at multiple locations and not only at the SGAGKT hexapeptide sequence at the end of the PA chain. The phosphate is observed to associate with the GGG linker, as well as between both GGG and SGAGKT regions; this suggests a rich potential sequence design space, in which linker sequences and phosphate-binding sequences may be screened for optimal phosphate sequestration.

Furthermore, adaptive biasing force simulations of a single C16GGGhex chain binding with phosphate were carried out at low and high pH conditions. Preferential binding is observed to occur at low pH conditions, with three distinct free energy minima identified along the peptide region of the PA. These minima are separated by moderate free energy minima, suggesting that phosphate, once bound, can move along the peptide region in a delocalized manner with intermittent hops between specific phosphate binding sites. ABF studies of the system at high pH indicate no preferential binding between C16GGGhex and phosphate, with a broad free energy minimum centered where the phosphate is unbound.

We then used ABF to study binding of a single phosphate to a flat layer of C16GGGhex at both high and low pH as a basic characterization of phosphate binding to an assembled C16GGGhex structure. At low pH, a free energy minimum of 39.7 kJ/mol per PA chain is observed to correspond to successful binding with phosphate; this well is broad, spanning 2 nm across, which is in agreement with our single-chain observation that phosphate binds to C16GGGhex in a delocalized manner. A small free energy mininum of 2.5 kJ/mol is observed for high pH conditions; taken together with the single-chain results at high pH, this suggests that phosphate may be captured mildly at high pH, but only in the presence of multiple, closely-packed C16GGGhex chains.

Although the work here lays a basic foundation for understanding a preliminary PA candidate for phosphate binding, our findings indicate multiple directions for future work and for improved design of phosphate-binding PAs. As discussed above, our observation that phosphate can bind with the linker in C16GGGhex, as well as between the linker and the SGAGKT phosphate-binding sequence, indicates that there is a rich sequence design space to be explored, where both the phosphate-binding sequence (based on the P-loop) and linker sequences may be screened in sequence space and tested for optimal phosphate capture; this work is currently underway in our laboratory, as well as studies extending our phosphate-binding screening to cylindrical micellar structures, which will shed light on binding behavior to a flexible self-assembled PA systems with increased surface area and accessible angles between adjacent PA chains.

4.4 Methods

We use a combination of GROMACS 5.1.4[2] and the ABF[26] enhanced sampling method as implemented in SSAGES[108] to study C16GGGhex binding to phosphate. C16GGGhex and phosphate at both pH conditions were modeled using the CHARMM force field,[10] and water was modeled using the TIP3P model.[52] For the brute force molecular dynamics simulations, a single C16GGGhex chain and a single phosphate ion were placed in a cubic box (side length 7 nm for low pH, and side length 10 nm for high pH) and then solvated, with temperature coupling set at 300 K and volume kept constant. At each pH condition, phosphate was initialized from 9 different starting configurations around the single C16GGGhex and MD was run for 20 ns from each of these positions, creating a total of 180 ns simulated at each pH. These trajectories were then analyzed by first rotationally and translationally aligning snapshots based on protein coordinates and then using the GROMACS `cluster` tool to cluster by protein backbone and phosphate coordinates in order to identify the most frequently occurring C16GGGhex-phosphate configurations.

Based on the multiple binding locations of phosphate on the peptide region of C16GGGhex, we pick the two distance CVs to characterize single-chain binding. First, we choose d_{SGAGKT}, which measures separation between the phosphate and the center of the SGAGKT binding pocket; this is defined by the distance between the center of mass of all phosphate atoms and the center of mass of the following atoms: {Backbone N on 8GLY, backbone N on 9LYS, sidechain N on 9LYS, backbone N on 10THR, sidechain O on 10THR}. Second, we choose d_{GGG}, which measures separation between the phosphate and the center of the GGG linker region; this is defined by the distance between the center of mass of all phosphate atoms and the center of mass of the following atoms: {Backbone N on 2GLY, backbone N on 3GLY, O on 3GLY, backbone N on 5SER, and sidechain O on 5SER}.

2-dimensional ABF simulations for single-chain binding to phosphate were carried out using d_{SGAGKT} and d_{GGG} as CVs, with bounds of [0.05 nm, 2.0 nm] for each CV and 50 bins are used for each CV. 4 walkers were used for each ABF run. Restraints were placed for each CV at values of 0.0 nm and 2.5 nm with a spring constant of 500 kJ/mol nm^2 in order to ensure that the configurations explored remained in the CV space of interest. Minimum visits to each ABF bin before forces are estimated is kept at 400. ABF was carried out and output monitored at intervals of 10 ns until the root mean squared error compared to the most recent output reached a plateau, resulting in 160 ns total simulation time per walker for the high pH system and 250 ns per walker for the low pH system.

ABF simulations for the flat C16GGGhex layer were carried out using distance between the z-coordinate of the center of mass of the phosphate and the z-coordinate of the center of mass of the tail-end alkyl carbons, with 50 bins used across the bounds of [0.05 nm, 9.0 nm]. 4 walkers were used for each ABF run. Restraints were placed at 0 nm and 9.5 nm with a spring constant of 500 kJ/mol nm^2, and minimum visits for force estimates again kept at 400. ABF was carried out and monitored at intervals of 10 ns for a total of 60 ns per walker for both high and low pH systems.

CHAPTER 5

EXTRACTING COLLECTIVE MOTIONS UNDERLYING NUCLEOSOME DYNAMICS VIA NONLINEAR MANIFOLD LEARNING

The identification of effective collective variables remains a challenge in molecular simulations of complex systems. Here, we use a nonlinear manifold learning technique known as the diffusion map to extract key dynamical motions from a complex biomolecular system known as the nucleosome: a DNA-protein complex consisting of red a DNA segment wrapped around a disc-shaped group of eight histone proteins. We show that without any *a priori* information, diffusion maps can identify and extract meaningful collective variables that characterize the motion of the nucleosome complex. We find excellent agreement between the collective variables identified by the diffusion map and those obtained manually using a free energy-based analysis. Notably, diffusion maps are shown to also identify subtle features of nucleosome dynamics that did not appear in those manually specified collective variables. For example, diffusion maps identify the importance of looped conformations in which DNA bulges away from the histone complex that are important for the motion of DNA around the nucleosome. This work demonstrates that diffusion maps can be a promising tool for analyzing very large molecular systems and for identifying their characteristic slow modes. This chapter is reproduced from [39].

5.1 Introduction

The continued development of advanced sampling techniques has extended the reach of molecular simulations considerably, thereby enabling the study of molecular systems of substantial complexity.[107] Higher complexity is accompanied by the challenge of describing key molecular processes. Ideally, we desire for these complex dynamics to be represented by a few low-dimensional descriptors, but automatically identifying such descriptors and quantifying how well they capture the system's dynamics can be challenging.

A range of approaches is available to discover these low-dimensional descriptors from simulated trajectories of a particular system. One attractive option is to use a dimensionality reduction technique to furnish a low-dimensional embedding of data from molecular dynamics trajectories,[30] using algorithms such as principal component analysis (PCA[51]), isometric feature map (Isomap[112]), locally linear embedding (LLE [97]), sketch-maps,[17] and diffusion maps.[22, 23] Diffusion maps have been widely applied to a variety of molecular systems, including all-atom miniprotein folding, [54] self-assembly of patchy colloids, [67] and coarse-grained protein models. [96] Furthermore, they have been adopted as part of multiple accelerated sampling algorithms, such as diffusion-map-directed MD (DM-d-MD[124]) and intrinsic map dynamics (iMapD[19]), and variations on the diffusion map itself have also been developed in order to address challenges in working with data with inhomogeneous densities and to reduce computational costs. [96, 116]

While diffusion maps have been applied in diverse contexts, there remain interesting challenges in applying diffusion maps to large and complex macromolecular systems, which exhibit inherently rich dynamics. One such system is the nucleosome, a DNA-protein complex consisting of a DNA segment wrapped around a disc-shaped complex of eight histone proteins.[71] The nucleosome is the basic building block of eukaryotic chromatin, which packs into successively higher-order structures in order to form the mitotic chromosome. Nucleosome positions and proper packaging of DNA are important for healthy cellular function.[42, 7]

Recent work has shown that DNA sequence is a key factor that governs nucleosome position, with different DNA sequences exhibiting different affinities for the histone octamer. The probability of nucleosome formation changes with this affinity and can span orders of magnitude across different DNA sequences. Several studies on DNA repositioning have been carried out, leading to the identification of two major repositioning mechanisms: (1) the loop propagation model,[101, 56, 68, 109, 93, 87] in which a loop of DNA is formed on one side of the nucleosome and moves in an inchworm-like manner along the histone complex; and (2) the twist diffusion model,[34, 110, 94, 57] in which a twist defect is introduced into the natural helicity of the DNA and diffuses in a corkscrew-like manner along the histone complex. Recent work by Lequieu et al.[60] investigated the relationship between DNA sequence and repositioning dynamics using a molecular model of the nucleosome; that study showed that different DNA sequences indeed rely on different mechanisms to reposition through pathways reminiscent of the proposed looping and twisting processes.

The simulations performed to reach these conclusions were considerably demanding, and required over 5 microseconds of unbiased simulation data, for 9 different DNA sequences. In the study of Lequieu et al. however, the order parameters used to characterize DNA motion were identified manually, and were necessarily influenced by human biases. As such, it is unclear if they can fully represent the true underlying dynamics of the nucleosome. The order parameters used in Lequieu et al. were based on the two previously proposed repositioning mechanisms, and thus analysis of the simulations focused specifically on loop propagation and twist diffusion. It is conceivable that other motions within the nucleosome might play important roles in cellular function, and may have been overlooked in this prior study.

In this work, we exploit this wealth of molecular dynamics data to interrogate the dynamics of the nucleosome using the diffusion map. This approach represents a bias-free method for identifying the collective variables that dominate nucleosomal motions. We show that a diffusion map approach is effective for identifying the collective variables previously found by Lequieu et al. through a detailed free energy analysis. Notably, without any *a priori* information, the diffusion map can distinguish DNA sequences that reposition via loop propagation from those that reposition via twist diffusion. Furthermore, the diffusion map approach is able to identify subtle molecular motions involving looping conformations, in which DNA bulges away from the histone octamer, and DNA breathing, in which DNA spontaneously unwraps from the histone complex. By applying the diffusion map to nucleosome dynamics, we show that both dominant and subtle dynamical modes can be automatically extracted from molecular simulation data, thereby reinforcing the diffusion map as a useful tool for unraveling the behavior of complex biomolecular systems.

5.2 Methods

5.2.1 MD Simulations of the Nucleosome

Molecular dynamics simulations were carried out with in-house codes using a coarse-grained representation of the 223 base pair nucleosome, as described in Lequieu et al.[60] The 3SPN.2C model is used to represent DNA and is the most recent version of the 3SPN model,[55, 98, 45, 33] in which DNA is represented by three sites at the centers of mass of phosphate, sugar, and base of each DNA nucleotide. The 3SPN.2C model has been further parameterized to capture the correct melting behavior of double-stranded to single-stranded DNA, sequence effects, and salt effects. The AICG model is used to represent the histone proteins, using a single site per amino acid at the side chain center of mass.[62] Interactions between the DNA and histone proteins consist of excluded volume effects and electrostatic forces, calculated using Debye-Hückel theory. Molecular dynamics simulations were performed in the canonical ensemble using a Langevin thermostat and ionic strength of 150 mM, with frames saved for later analysis every 1 ns. Further details can be found in Lequieu at al.[60]

5.2.2 Diffusion maps

Diffusion maps are a type of nonlinear dimensionality reduction technique originally introduced by Coifman and co-workers.[22, 23] Here, we briefly step through the algorithm to clarify and facilitate subsequent discussion. Specifically, we use the density-adapted diffusion map introduced by Wang and Ferguson[116] due to the inhomogeneous sampling of configurations in brute-force molecular dynamics simulations of the nucleosome.

First, pairwise distances d_{ij} are calculated between datapoints $\mathbf{x}_i$ and $\mathbf{x}_j$. In this case, we use the root-mean-squared distance between translationally and rotationally aligned atomic coordinates between two molecular configurations. d_{ij} is then passed through a Gaussian kernel to construct matrix $\mathbf{A}$, which contains the now thresholded pairwise distances, with entries

$$A_{ij} = \exp\left(\frac{-d_{ij}^{2\alpha}}{2\epsilon}\right). \tag{5.1}$$

Here, ϵ is the kernel bandwidth and α rescales pairwise distances globally in order to smooth out density inhomogeneities in sampled configurations. We find that an α value of 0.3 works well for configurations from all three DNA sequences considered here. The kernel bandwidth ϵ defines the extent of the local neighborhood around each datapoint in which to consider pairwise distances to other points, and we use an ϵ of 3.0 for our data across all sequences. $\mathbf{A}$ is then row-normalized to form the Markov matrix

$$\mathbf{M} = \mathbf{D}^{-1}\mathbf{A}, \tag{5.2}$$

where $\mathbf{D}$ is a diagonal matrix with entries

$$D_{ij} = \sum_j A_{ij}. \tag{5.3}$$

$\mathbf{M}$ is effectively a transition matrix, with entries M_{ij} corresponding to transition probabilities between configurations $\mathbf{x}_i$ and $\mathbf{x}_j$.

Finally, $\mathbf{M}$ is diagonalized in order to calculate its eigenvectors $\{\psi_i\}$ and associated eigenvalues $\{\lambda_i\}$. Due to the Markovian nature of $\mathbf{M}$, the top eigenvalue-eigenvector pair (ψ_0, λ_0) is trivial; this pair corresponds to the steady-state distribution of a random walk with $\lambda_0 = 1$.

By locating a gap in the eigenvalue spectrum between λ_k and λ_{k+1}, one can identify the top k non-trivial eigenvectors $\{\psi_i\}_{i=1}^{k}$ corresponding to slow diffusion modes of the system, which dominate over the fast modes corresponding to the remaining lower eigenvectors $\{\psi_i\}_{i>k}$. The original high-dimensional data can then be embedded in k dimensions by projecting the data onto the top k non-trivial eigenvectors,

$$\mathbf{x}_i \mapsto [\psi_1(i), \psi_2(i), \ldots, \psi_k(i)]\,. \tag{5.4}$$

In some cases, multiple gaps may emerge in the eigenvalue spectrum, in which case one must avoid only using eigenvectors up to the first gap, which may produce misleading results. The final low-dimensional embedding reflects the intrinsic manifold underlying the molecular system as extracted from the sampled molecular dynamics data.

Analysis of nucleosome simulations using the density-adapted diffusion map began with calculation of $\mathbf{A}$ for each DNA sequence studied, using Equation 5.1 and snapshots extracted from MD simulation trajectories. $\mathbf{M}$ was then calculated for each sequence as described above in Equations 5.2 and 5.3, followed by calculation of eigenvectors $\{\psi_i\}$ and eigenvalues $\{\lambda_i\}$ for each sequence's $\mathbf{M}$. The spectra of $\{\lambda_i\}$ were examined visually in order to identify gaps and determine non-trivial eigenvectors for each sequence-specific diffusion map embedding. Multiple collective variables (described in the following three subsections) were calculated for each simulation snapshot used to create the embeddings and then projected onto the non-trivial eigenvectors to create diffusion map embeddings of collective variables for each sequence. These diffusion map embeddings of collective variables were then used to identify sequence-specific correlations of collective variables with dominant dynamical modes of the nucleosome system.

5.2.3 Collective Variables Describing DNA Translocation and Rotation

DNA translocation relative to the histone dyad is characterized by S_T, defined as:

$$S_T = \left\langle \pm \arccos\left(\frac{\mathbf{P}\cdot\mathbf{P_0}}{\|\mathbf{P}\|\|\mathbf{P_0}\|}\right)\right\rangle. \tag{5.5}$$

Here, vector $\mathbf{P}$ points from the center of a base step to the center of the protein complex, and $\mathbf{P_0}$ is the corresponding value of $\mathbf{P}$ taken from a reference nucleosome crystal structure (PDB ID: 1KX5),[28] which was used to create initial structures for the nucleosome simulations. The average in Equation 5.5 is taken over -15, -5, +5, and +15 base steps relative to the histone dyad, located at the central position on the nucleosome (indicated by the triangle in Figure 5.1). If $(\mathbf{P} \times \mathbf{P_0}) \cdot \hat{\mathbf{f}} \leq 0$ then the positive sign is used (otherwise, negative), where vector $\hat{\mathbf{f}}$ points along the center of the nucleosomal DNA superhelix. Using this sign convention, positive S_T corresponds to forward translocation of DNA toward the 5' end, while negative S_T corresponds to reverse translocation toward the 3' end.

A second nucleosome repositioning order parameter is S_R, which characterizes DNA rotation:

$$S_R = \left\langle \pm \arccos\left(\frac{\mathbf{P} \cdot \mathbf{B}}{\|\mathbf{P}\|\|\mathbf{B}\|}\right)\right\rangle, \tag{5.6}$$

where vector $\mathbf{B}$ points from the center of a given base step on the sense strand to its complementary base step on the anti-sense strand. $\mathbf{P}$ and the average denoted by the angle brackets are as defined for S_T. If $(\mathbf{P} \times \mathbf{B}) \cdot \mathbf{D} \leq 0$, then the positive sign is used (otherwise, negative). $\mathbf{D}$ is a vector in the 5' to 3' direction along the sense strand of the DNA. If $S_R = -\frac{\pi}{2}$, the minor groove of the DNA double helix is oriented toward the histone core, whereas when $S_R = \frac{\pi}{2}$, the minor groove is oriented away from the histone complex.

5.2.4 Collective Variable Describing DNA Breathing

DNA breathing, which involves spontaneous unwrapping and rewrapping of DNA from the nucleosome, was characterized by two angle parameters, θ_{forward} and θ_{backward}, shown in Figure 5.2. Each angle is calculated between a vector from the center of mass of the histone to the dyad, which is relatively immobile, and a vector from the 30th DNA base pair to the first and endmost DNA base pair, which moves significantly as DNA unwraps.

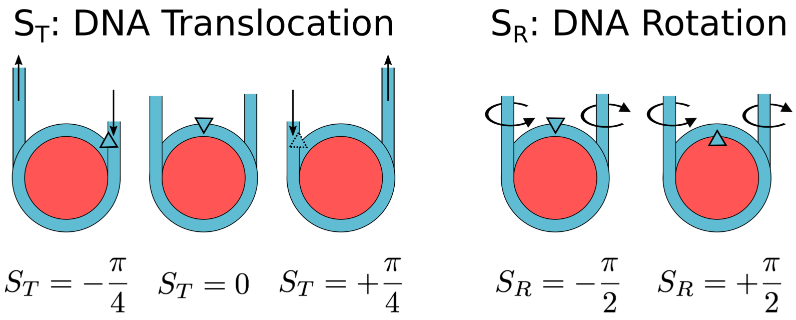


Figure 5.1: Schematic of order parameters S_T and S_R, which characterize DNA translocation and DNA rotation, respectively. These order parameters are defined in Section 5.2.3.

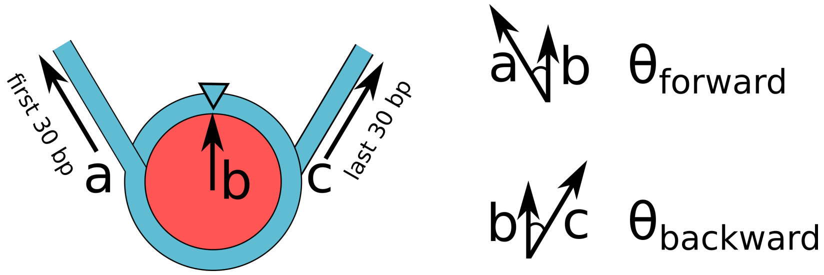
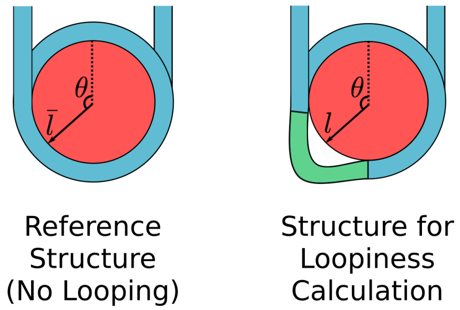


Figure 5.2: Schematic of order parameters θ_{forward} and θ_{backward} characterizing DNA breathing, in which strands of DNA spontaneously unwrap and rewrap from the histone complex. Each angle is calculated from two vectors: vector **b** points from the histone complex center of mass to the dyad, and vectors **a** and **c** point along either end of the DNA strand, from the 30th base pair to the first and endmost base pair.


θ
l̄
Reference Structure (No Looping)
θ
l
Structure for Loopiness Calculation

$$\Delta l(\theta) = \langle l(\theta) \rangle - \langle \bar{l}(\theta) \rangle$$
$$\Delta l^*(\theta) = \max(\Delta l(\theta) - 8\text{Å}, 0)$$
$$\text{loopiness} = \int_\theta \Delta l^*(\theta) d\theta$$

Figure 5.3: Schematic of the loopiness order parameter, which characterizes the extent to which DNA bulges away from the histone octamer. Calculation of this order parameter is described in Section 5.2.5.

5.2.5 Collective Variable Describing DNA Looping

DNA looping, which involves DNA bulging away from the histone octamer, was characterized using a *loopiness* order parameter (Figure 5.3). To calculate loopiness, we first calculate two values for each ith DNA base pair: the distance of the base pair to the histone center of mass, l_i, and the location of the base pair relative to the dyad, θ_i. For ease of calculation, we compute the average distance from a base pair to the histone center of mass as a function of location θ, denoted as $\langle l(\theta) \rangle$. In order to normalize $\langle l(\theta) \rangle$, we then calculate the corresponding value of this average distance in the complete absence of looping, $\langle \bar{l}(\theta) \rangle$, which is calculated from a nucleosome simulation performed in very low salt concentration for a strongly binding DNA sequence. We then normalize $\langle l(\theta) \rangle$ using $\langle \bar{l}(\theta) \rangle$ by calculating deviation from the loop-free case $\Delta l(\theta) = \langle l(\theta) \rangle - \langle \bar{l}(\theta) \rangle$; in cases where there is no DNA looping, Δl is approximately 0 across all locations θ, and in cases where DNA loops form, $\Delta l > 0$. To eliminate baseline noise, we threshold Δl by subtracting a threshold value of 8Å, which corresponds to the Debye length at 150 mM at which DNA-histone attraction has largely decayed. The post-threshold looping parameter Δl^* is then integrated along the entire circumference around the histone octamer (over all θ) in order to obtain our final *loopiness* order parameter.

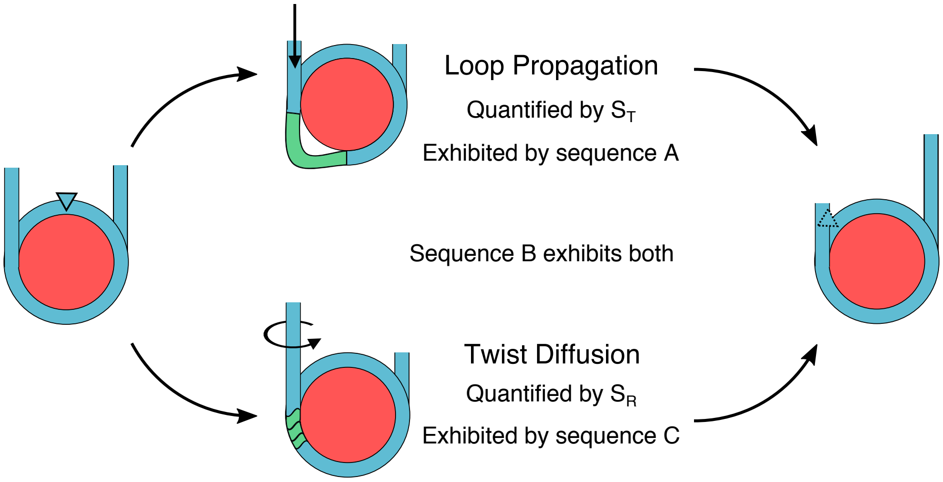


Figure 5.4: Schematic showing two proposed nucleosome repositioning mechanisms: loop propagation and twist diffusion. The histone complex is represented in red, and DNA in blue. S_T and S_R quantify loop propagation and twist diffusion, respectively; definitions for these order parameters are introduced in the Methods section. Individual repositioning propensities for sequences A, B, and C are also shown.

5.3 Results and Discussion

We apply the diffusion map to a subset of these trajectories from three representative DNA sequences: sequence A, a strongly binding sequence that primarily repositions by loop propagation; sequence B, a moderately binding sequence that exhibits a combination of loop propagation and twisting; and sequence C, a weakly binding sequence that primarily repositions by twisting. These sequences are tabulated in Table 5.1 with their respective binding strengths and sequence identities. Figure 5.4 summarizes the loop propagation and twisting models of nucleosome repositioning, along with the respective repositioning behaviors for all three sequences studied and the collective variables used to describe repositioning, which will be introduced later in the text.

By applying the density-adapted diffusion map on configurations for sequences A, B, and C as described in the previous section, we obtain the eigenvalue spectra shown in Figure 5.5. Snapshots for the diffusion map analysis were extracted from the molecular dynamics trajectories at evenly spaced intervals (every 40 ns for sequences A and B, and every 25 ns for sequence C), for a total of 16,207 snapshots from sequence A, 14,917 from sequence

Table 5.1: DNA sequences used in this work, along with their binding strengths and sequence names used in the literature.

Sequence Name	Binding Strength	Name in Literature
A	Strong	c3 (See [102])
B	Moderate	TRGC (See [81] and [80])
C	Weak	TTAGGG (See [106])

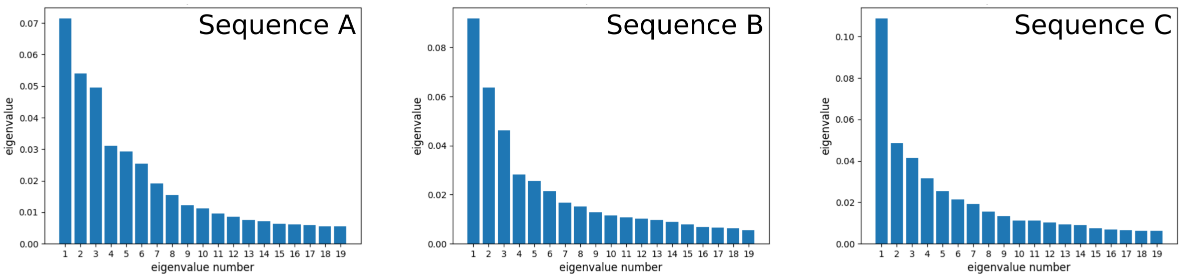


Figure 5.5: Eigenvalue spectra for sequences A, B, and C. Note that sequences A and B exhibit hierarchical character, indicated by multiple gaps in the eigenvalue spectra. Both spectra show three dominant eigenvalues, followed by three moderate eigenvalues, suggesting that three major slow dynamical modes dominate the system, while three less significant modes still contribute to the system dynamics.

B, and 10,713 from sequence C. Sequences A and B exhibit similar hierarchical eigenvalue spectra, indicated by multiple spectral gaps. Both sequences exhibit gaps between ψ_3 and ψ_4, and between ψ_6 and ψ_7, suggesting that dynamics are dominated by a combination of three major slow modes (ψ_1 to ψ_3) and three moderate modes (ψ_4 to ψ_6). The eigenvalue spectrum for sequence C exhibits a large, distinct gap after ψ_1 and a smaller gap after ψ_3, which indicates that one major slow mode dominates the system, followed by two moderate modes.

5.3.1 DNA Translocation

First, we check if the diffusion map is able to recover the two nucleosome repositioning order parameters studied in Lequieu et al.[60] We begin with S_T, the order parameter characterizing DNA translocation relative to the histone dyad. Figure 5.6 shows two- and three-

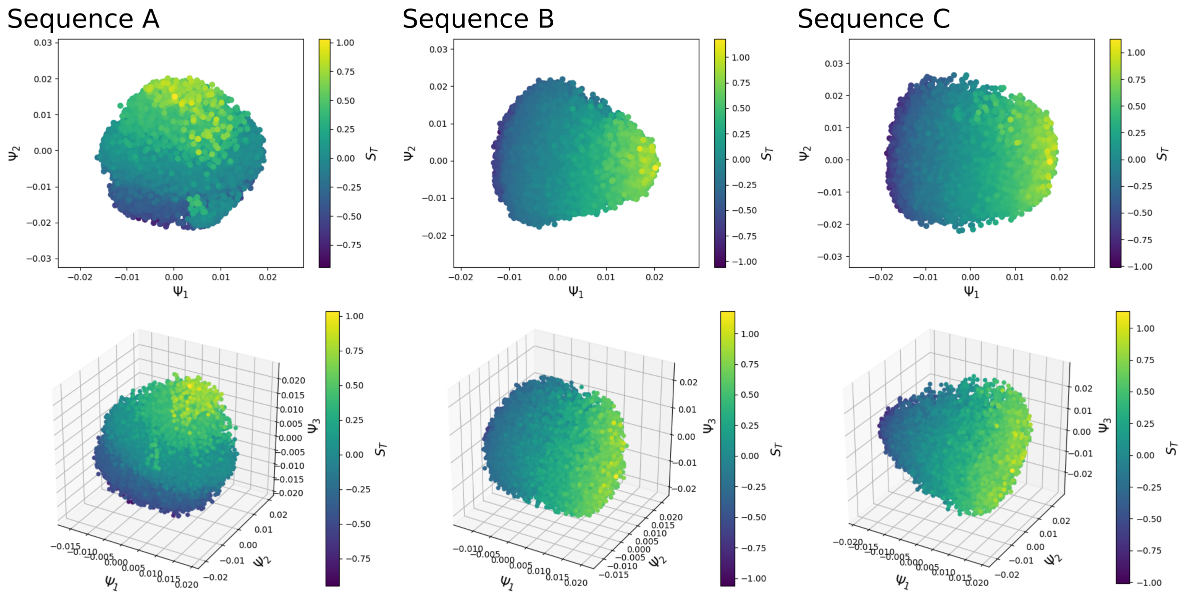


Figure 5.6: 2- and 3-dimensional diffusion map embeddings of S_T for all sequences. DNA translocation correlates with ψ_2 for sequence A, indicated by the gradient in S_T along the vertical ψ_2 axis. DNA translocation correlates with ψ_1 for sequences B and C, indicated by the gradient in S_T along the horizontal ψ_1 axis.

dimensional diffusion map embeddings for all three sequences studied, using the first three non-trivial eigenvectors and colored by S_T. In all three sequences, DNA translocation is found to be well parameterized by either the slowest (ψ_1) or second slowest (ψ_2) dynamical mode identified by the diffusion map, indicating that S_T correlates with slow modes across binding affinities. The correlation of S_T with either ψ_1 or ψ_2 in all three sequences supports the idea that there will always be some degree of translocational motion in the nucleosome repositioning process, regardless of the preference for a particular DNA sequence to reposition by either looping or twisting.

5.3.2 DNA Rotation

Figure 5.7 shows diffusion map embeddings of S_R, which quantifies DNA rotation, for all three sequences studied, using the top three non-trivial eigenvectors and colored by S_R. There is no correlation of S_R with these top three eigenvectors for sequences A and B; further analysis confirms that S_R is not well parameterized by any of the top six eigenvectors for these sequences. This is expected, since A and B exhibit relatively strong binding affinities and are more likely to reposition by a looping mechanism as opposed to a twisting mechanism.

In contrast, sequence C, a weakly binding sequence that primarily repositions by rotation, exhibits a periodic banded structure, which appears more clearly in the two-dimensional embedding of sequence C in ψ_1 and ψ_3 (Figure 5.7d). Furthermore, we can construct an effective free energy landscape from the diffusion map embedding of sequence C by collecting a histogram of sequence C datapoints in ψ_1 and S_R, normalizing by the total number of datapoints so that the resulting probability in the bins sum to 1, and taking the negative logarithm of these probabilities. This effective free energy landscape is plotted in Figure 5.7e; this is reminiscent of the free energy landscape calculated for sequence C using conventional methods (umbrella sampling and WHAM) found by Lequieu et al.,[60] plotted in S_T vs S_R and reproduced in Figure 5.7f; this is consistent with our earlier finding that S_T correlates with ψ_1 for this sequence.

The order parameters characterizing DNA translocation and rotation emerge in the same non-trivial eigenvector for sequence C, consistent with prior observations that sequence C repositions via twisting. In contrast, only translocation is extracted from the underlying MD data for sequences A and B, consistent with prior observations that sequences A and B do not reposition through DNA twisting. Through analysis of all three sequences, we observe that the diffusion map approach identifies a slow mode that correlates with DNA translocation across all binding strengths. DNA rotation emerges in the same slow mode if the sequence exhibits repositioning by rotation as well, suggesting that this particular non-trivial eigenvector corresponds to a more general repositioning motion consisting of a combination of translocation and, if the sequence exhibits it, rotation.

5.3.3 DNA Breathing

Next, we examine whether the diffusion map approach can be used to identify key nucleosome dynamics beyond the translocational and rotational repositioning mechanisms studied

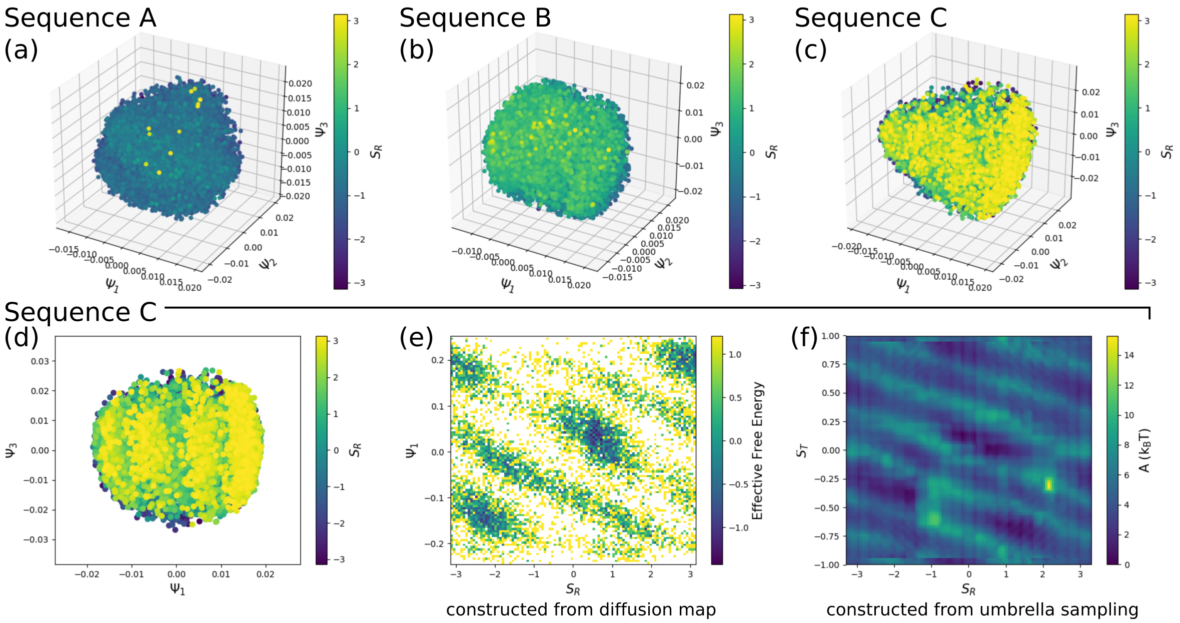


Figure 5.7: (a-c): 3-dimensional diffusion map embeddings of S_R, the order parameter that characterizes DNA rotation, for all sequences. For clarity, datapoints with greater values of S_R are shown at higher layers of the plot. There is no correlation of S_R with top non-trivial eigenvectors for sequences A and B; (d) 2-dimensional diffusion map embedding of S_R for sequence C. ψ_1 correlates with cycles of DNA rotation, as indicated by the periodic bands of S_R along ψ_1; (e) effective free energy constructed from the diffusion map embedding for sequence C. Effective free energy is calculated by histogramming datapoints for sequence C in ψ_1 and S_R, normalizing each histogram bin by the total number of datapoints to calculate probabilities, and then taking the negative log of each bin. The resulting density plot exhibits a periodic banded structure reminiscent of the free energy landscape for sequence C constructed by conventional methods by Lequieu et al.,[60] which is plotted in (f) in S_T vs S_R. Note that S_T was previously found to correlate with ψ_1 for sequence C; the diffusion map has effectively unfurled the same previously calculated free energy landscape.

in Lequieu et al.[60] One particularly interesting aspect of nucleosome dynamics is DNA breathing, which involves unwrapping of nucleosomal DNA from the histone complex. Single-molecule FRET experiments have shown that nucleosomal DNA can spontaneously unwrap and rewrap from the histone octamer, allowing transcription factors, enzymes, and other proteins to interact with previously unaccessible portions of DNA that were buried by the histone complex.[61, 113]

Figure 5.8 shows two-dimensional diffusion map embeddings for all three sequences, colored by the average of θ_{forward} and θ_{backward}, which captures breathing on both sides of the nucleosome. The average breathing order parameter correlates with ψ_2 for sequence A, and with ψ_1 for sequences B and C. Interestingly, in each sequence, the average breathing order parameter correlates with the same eigenvector as S_T (and S_R, in the case of sequence C); this is evident in the visual similarities between Figures 5.6 and 5.8. The shared correlations of the average breathing order parameter with S_T and S_R suggest that repositioning dynamics and breathing dynamics are closely tied. The embeddings generated by the diffusion map approach capture both of these motions within the same non-trivial eigenvector, implying that these two types of dynamics are innately part of the same characteristic dynamic mode exhibited by the nucleosome. Although the diffusion map is unable to provide an explicit nonlinear mapping from the high-dimensional input to low-dimensional coordinates, and interpretation of the low-dimensional coordinates is limited to correlating the top eigenvectors of $\mathbf{M}$ with various descriptors of the system, this perceived deficiency may also be interpreted as an advantage, since it provides a tool for identifying multiple CVs that may be coupled together in the same slow dynamical mode, as we have just observed with S_T and S_R for sequence C.

5.3.4 DNA Looping

In Figure 5.5, sequences A and B were found to exhibit hierarchical eigenvalue spectra, with three dominant non-trivial eigenvectors (ψ_1, ψ_2, ψ_3) and three moderate non-trivial

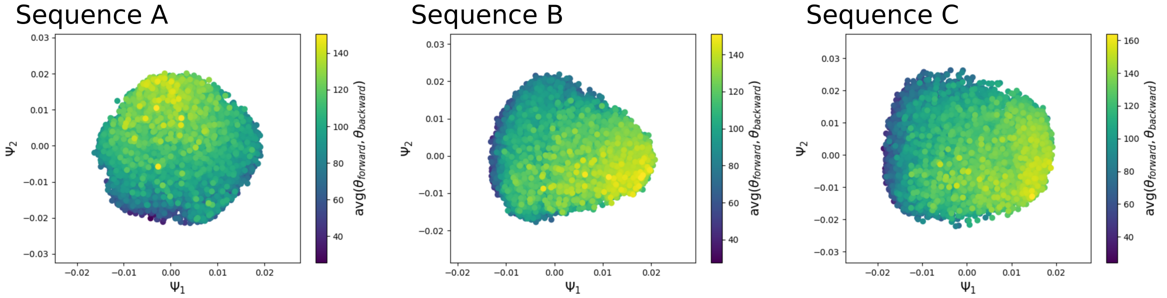


Figure 5.8: Two-dimensional embeddings of the average of θ_{forward} and θ_{backward}, which characterizes DNA breathing, for all sequences. The average breathing order parameter for sequence A correlates with ψ_2, as indicated by the gradient in the breathing order parameter along the vertical ψ_2 axis; ψ_2 also correlates with the order parameter characterizing DNA translocation, S_T, for sequence A, as seen in Figure 5.6. The average breathing order parameter for sequences B and C correlate with ψ_1, as indicated by the gradient in the breathing order parameter along the horizontal ψ_1 axis; this eigenvector also correlates with S_T for these two sequences, again as seen in Figure 5.6. For sequence C, this eigenvector also correlates with S_R, as seen in Figure 5.7.

eigenvectors (ψ_4, ψ_5, ψ_6). Our analysis thus far, using the diffusion map approach, has focused on motions correlating with the top group of non-trivial eigenvectors. We now examine the significance of the moderate non-trivial eigenvectors in sequences A and B (and why this feature is absent from the eigenvalue spectrum for sequence C).

Figure 5.9 shows two-dimensional diffusion map embeddings of the loopiness order parameter, described in the Methods section, for sequences A, B, and C using the moderate eigenvectors ψ_4, ψ_5, and ψ_6. Protrusions are observed in all three embeddings for sequence A and the embedding of sequence B in ψ_5 and ψ_6. Through visual inspection of configurations corresponding to points within and outside of the protruding lobe, we find that the protrusion corresponds to configurations exhibiting DNA loops. In more “loopy” configurations, DNA bulges away from the histone complex, and gaps are formed between the DNA and histone octamer. Loopy configurations are necessary for the loop propagation involved in DNA translocation characterized by the order parameter S_T, as described earlier, with translocation dominating in more strongly binding sequences.

The emergence of loopiness in ψ_4, ψ_5, and ψ_6 in sequences A and B is consistent with their relative propensities for translocation. For strongly binding sequence A, loopiness emerges

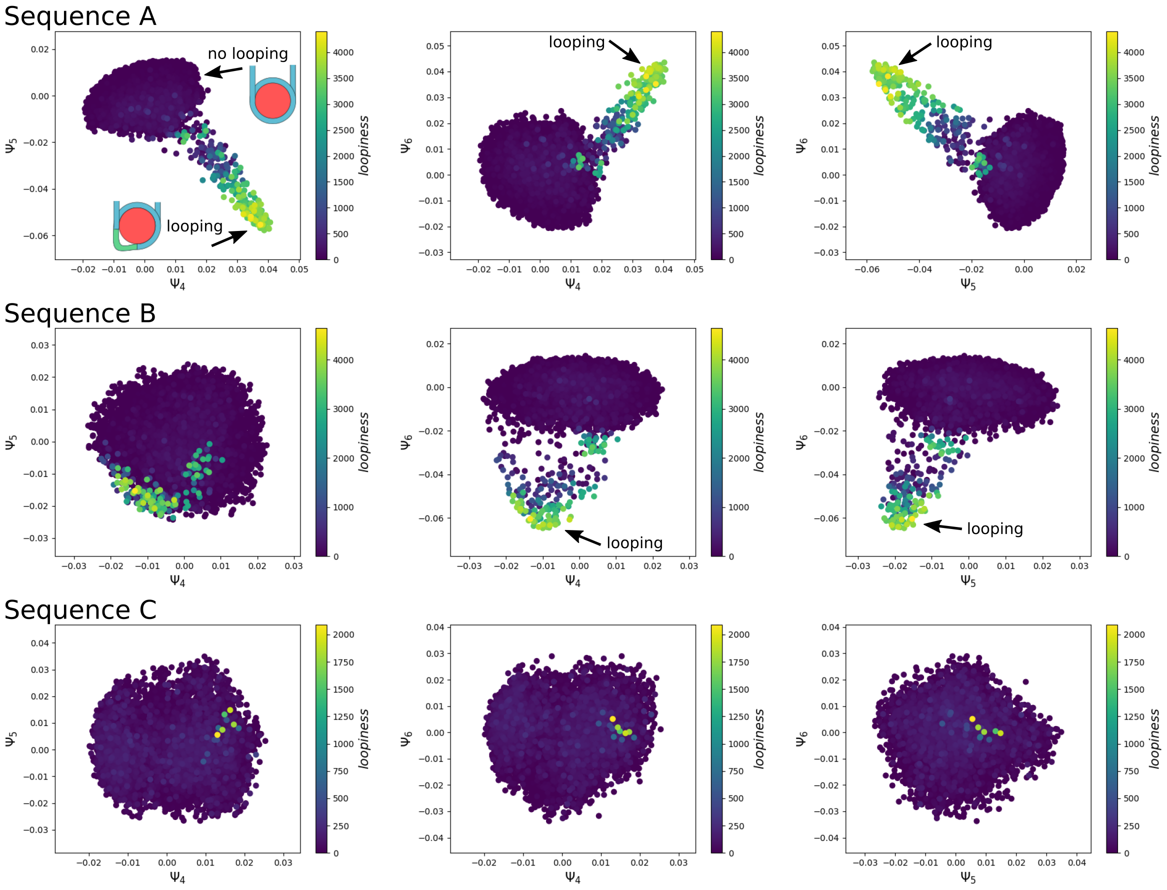


Figure 5.9: Two-dimensional diffusion map embeddings of loopiness for all sequences, plotted by moderate eigenvectors ψ_4, ψ_5 and ψ_6. For sequence A, more loopy configurations are isolated by all three moderate eigenvectors. For sequence B, loopy conformations are only isolated by ψ_6. Loopy configurations are not extracted for sequence C.

in multiple higher eigenvectors compared with moderately binding sequence B; loopy configurations for sequence A are clearly isolated in ψ_4 through ψ_6. In contrast, loopiness only emerges in ψ_6 for sequence B, which exhibits a lower propensity for translocation compared to sequence A. Furthermore, weakly binding sequence C repositions entirely by rotation and does not exhibit any moderate eigenvectors. In fact, loopiness does not emerge in any of the top 12 eigenvectors for sequence C.

DNA loop formation is important well beyond the context of the mechanics of loop propagation, with implications in chromatin remodeling and spontaneous nucleosome migration,[87] and we show that the diffusion map can automatically identify this subtle mode. Furthermore, we find that looping is embedded in higher-order eigenvectors, which diffusion map studies often bypass while focusing on the first several dominant eigenvectors. These top eigenvectors often extract the dynamic modes corresponding to collective variables more easily identified by hand, as the present study shows with S_T and S_R. We show that thorough examination of higher-order modes can provide valuable insight into more subtle dynamics of complex systems that may be easier for humans to miss.

5.4 Conclusions

Diffusion maps were used to extract key motions underlying nucleosome dynamics from MD trajectories of nucleosome repositioning for three representative DNA sequences, spanning different binding strengths (and consequently, different repositioning dynamics). Translocational and rotational motions, which had been previously identified through a detailed free energy analysis by Lequieu et al.,[60] were confirmed by the diffusion map approach. Translocational motions were found to correlate with dominant slow modes across the three DNA sequences examined here. Rotational motions were only found to emerge in the weakest binding sequence studied, emerging in the same slow mode that correlates with translocation.

In addition to finding the previously reported translocational and rotational order parameters, the diffusion map analysis was also used to extract DNA breathing and looping motions. Measures of DNA breathing, in which DNA spontaneously unwraps from and rewraps around the histone complex, were found to correlate with the same eigenvectors that correlate with DNA translocation and rotation, suggesting that DNA repositioning and DNA breathing are inherently part of the same dynamical mode. Sequences that exhibit DNA sliding were found to exhibit hierarchical eigenvalue spectra, with looping configurations isolated in the moderate eigenvectors corresponding to eigenvalues between the first and second spectral gap. The dominance of DNA sliding over twisting is further reflected in the order in which loopiness appears in these moderate eigenvectors. Weakly binding sequence C, which primarily repositions by twisting, neither exhibited a hierarchical eigenvalue spectra nor any eigenvectors that correlated with loopiness.

The diffusion map approach is particularly useful in enabling the discovery of key dynamical motions directly from MD data without defining *a priori* what exactly these motions might be. Although interpretation of dominant dynamical modes is aided by embedding user-specified order parameters in the diffusion map, as done in this work, these order parameters need not be supplied in order to calculate the non-trivial eigenvectors corresponding to these dominant modes, nor specially created in order to interpret a specific non-trivial eigenvector. For example, one might interpret a particular eigenvector by visually examining snapshots of the simulation drawn from different areas of the diffusion map, or use a generalized collective variable instead (ex. fit an eigenvector as a function of atomic coordinates from each simulation snapshot in the diffusion map). Considering the importance of sequence dependence in nucleosome dynamics, diffusion maps provide an attractive solution for rapid screening and identification of key dynamics across sequences in more complex scenarios, for example in higher order chromatin structures or comparing across mutated sequences. Even in the single nucleosome case studied in this work, there remain several significant eigenvectors for which the corresponding dynamics are unknown; we are actively working on elucidating these dynamics. More generally, this work emphasizes the possibilities of uncovering unintuitive properties in MD data that may be missed by more traditional approaches. Here, we are able to confirm both previously known and new order parameters using a small subset (and only 3 out of 9 total sequences) of the MD trajectories previously used in a detailed free energy analysis, attesting to the usefulness and efficiency of applying diffusion maps to previously simulated complex systems.

CHAPTER 6

CONCLUSIONS

In this work, we have applied the tools of molecular simulation and data-driven analysis to the study of multiple biological macromolecule systems, revealing the underlying dynamics and thermodynamics of protein aggregation, ion capture by peptide amphiphile assemblies, and nucleosomal repositioning dynamics. In Chapters 2 and 3, we found that aggregation of human amylin in water is a process uphill in free energy for both the formation of the dimer and trimer, with greater free energy barriers encountered in trimer formation. Using the finite temperature string method, intermediates were found in both dimerization and 3-chain trimerization that agree with previous experimental characterization of the early-stage amylin assembly process, with intermediate β-sheet formation in residues L12A13 and 20–29. Furthermore, we found 2+1 trimerization to be much less favorable compared to 3-chain trimerization, and a thorough comparison of H-bond formation, entropic interactions versus enthalpic interactions, and decomposition of potential energy have linked this discrepancy to the role of protein-water hydrogen bonds in each trimerization scenario. Additionally, string method calculations for trimerization indicated that 3-chain assembly begins from the dimer intermediate, which is then incorporated into the trimer intermediate, suggesting a possible stepwise aggregation mechanism, associated with moderate free energy barriers. Our findings on the importance of H-bonds in the aggregation process also motivate questions on whether aggregation in the presence of other proteins, with H-bond-forming molecules, or at higher salt concentration will drive the system toward more favorable aggregate formation. Interestingly, these additional components would add physiologically relevant components to the system that are currently missing from our model. Remaining questions also include whether further growth toward larger oligomers will remain uphill in free energy, and whether spontaneous fibril growth occurs past a certain oligomer size; these may be approached by extending the string method approach or by pivoting to a Markov State Modeling approach.

In the following chapter, we applied the ABF method in the study of peptide amphiphile

C16GGGhex to calculate free energy landscapes corresponding to its phosphate-binding behavior at two different pH conditions. We found preferential binding to occur at low pH, but not at high pH conditions. Furthermore, we observed C16GGGhex to bind with phosphate at multiple locations along its peptide region, including at the linker region rather than with the adjacent phosphate-binding hexapeptide. These patterns are observed for both single-chain binding simulations and studies of phosphate binding to a flat layer of C16GGGhex. Further studies are necessary to evaluate phosphate-binding behavior when C16GGGhex is able to assemble into long, worm-like micelles or combined with a separate hydrogel-forming peptide amphiphile, which more accurately represent the system structure if deployed in a functional material.

Finally, in Chapter 5, we demonstrated that the diffusion map approach is effective for identifying and extracting meaningful collective variables that characterize the motion of the nucleosome complex. The diffusion map was used to confirm discovery of previously identified collective variables via a painstaking free energy-based analysis, as well as identify more subtle features of nucleosomal dynamics that were not incorporated into those manually specified descriptors, including looped conformations, in which the DNA bulges out from the histone complex, and breathing motions, in which DNA on either end of the nucleosome spontaneously unwraps from the histone complex.

While this work lays a foundation for understanding a variety of biological systems, further study involves the investigation of larger systems and the targeting of longer timescales, as alluded to previously. Our findings in Chapter 5 suggest that valuable insights may be gleaned from existing molecular simulation data in order to inform the direction of this future work. Interestingly, the free energy-based analysis from which the previously identified collective variables were drawn in Chapter 5 take similar approaches to those used in Chapters 2 through 4, perhaps indicating the potential usefulness of taking a two-pronged approach in future studies of complex biological systems by continuing to examine dynamics and thermodynamics via free energy methods while simultaneously mining the generated molecular simulation data. Taken together with the current advances in free energy methods that incorporate machine learning techniques,[40, 105] the future holds great opportunities that incorporate existing methodologies and data-driven approaches together for the computational study of complex biological systems and the novel engineered materials that they inspire.

REFERENCES

[1] Andisheh Abedini, Fanling Meng, and Daniel P. Raleigh. A single-point mutation converts the highly amyloidogenic human islet amyloid polypeptide into a potent fibrillization inhibitor. *J. Am. Chem. Soc.*, 129(37):11300–11301, 2007.

[2] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilrd Pll, Jeremy C. Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19 – 25, 2015.

[3] Handan Acar, Samanvaya Srivastava, Eun Ji Chung, Mathew R. Schnorenberg, John C. Barrett, James L. LaBelle, and Matthew Tirrell. Self-assembling peptide-based building blocks in medical applications. *Advanced Drug Delivery Reviews*, 110-111:65 – 79, 2017. Peptides and Peptide Conjugates in Medicine.

[4] Sahar Bedrood, Yiyu Li, J. Mario Isas, Balachandra G. Hegde, Ulrich Baxa, Ian S. Haworth, and Ralf Langen. Fibril structure of human islet amyloid polypeptide. *J. Biol. Chem.*, 287(8):5235–5241, 2012.

[5] Herman J. C. Berendsen, David van der Spoel, and R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.*, 91:43–56, 1995.

[6] Workalemahu Mikre Berhanu and Artëm E. Masunov. Full length amylin oligomer aggregation: insights from molecular dynamics simulations and implications for design of aggregation inhibitors. *Journal of Biomolecular Structure and Dynamics*, 32(10):1651–1669, 2014.

[7] Sukesh R. Bhaumik, Edwin Smith, and Ali Shilatifard. Covalent modifications of histones during development and disease pathogenesis. *Nat. Struct. Mol. Biol.*, 14(11):1008–1016, 2007.

[8] Antonio Bianchi, Claudia Giorgi, Paolo Ruzza, Claudio Toniolo, and E. James Milner-White. A synthetic hexapeptide designed to resemble a proteinaceous p-loop nest is shown to bind inorganic phosphate. *Proteins: Structure, Function, and Bioinformatics*, 80(5):1418–1424, 2012.

[9] Xevi Biarnés, Fabio Pietrucci, Fabrizio Marinelli, and Alessandro Laio. METAGUI. A VMD interface for analyzing metadynamics and molecular dynamics. *Comput. Phys. Commun.*, 183(1):203–211, 2012.

[10] Pr Bjelkmar, Per Larsson, Michel A. Cuendet, Berk Hess, and Erik Lindahl. Implementation of the charmm force field in gromacs: Analysis of protein stability effects from correction maps, virtual interaction sites, and water models. *Journal of Chemical Theory and Computation*, 6(2):459–466, 2010.

[11] Matthew Black, Amanda Trent, Yulia Kostenko, Joseph Saeyong Lee, Colleen Olive, and Matthew Tirrell. Self-assembled peptide amphiphile micelles containing a cytotoxic t-cell epitope promote a protective immune response in vivo. *Advanced Materials*, 24(28):3845–3849, 2012.

[12] Massimiliano Bonomi, Davide Branduardi, Giovanni Bussi, Carlo Camilloni, Davide Provasi, Paolo Raiteri, Davide Donadio, Fabrizio Marinelli, Fabio Pietrucci, Ricardo A Broglia, and Michele Parrinello. PLUMED: A portable plugin for free-energy calculations with molecular. *Comput. Phys. Commun.*, 180(10):1961–1972, 2009.

[13] Jeffrey R. Brender, Edgar L. Lee, Marchello A. Cavitt, Ari Gafni, Duncan G. Steel, and Ayyalusamy Ramamoorthy. Amyloid fiber formation and membrane disruption are separate processes localized in two distinct regions of IAPP, the type-2-diabetes-related peptide. *J. Am. Chem. Soc.*, 130(20):6424–6429, 2008.

[14] L. E. Buchanan, E. B. Dunkelberger, H. Q. Tran, P.-N. Cheng, C.-C. Chiu, P. Cao, D. P. Raleigh, J. J. de Pablo, J. S. Nowick, and M. T. Zanni. Mechanism of IAPP amyloid fibril formation involves an intermediate with a transient β-sheet. *Proc. Natl. Acad. Sci.*, 110(48):19285–19290, 2013.

[15] Steve R. Bull, Mustafa O. Guler, Rafael E. Bras, Thomas J. Meade, and Samuel I. Stupp. Self-assembled peptide amphiphile nanofibers conjugated to mri contrast agents. *Nano Letters*, 5(1):1–4, 2005.

[16] Giovanni Bussi, Alessandro Laio, and Michele Parrinello. Equilibrium Free Energies from Nonequilibrium Metadynamics. *Phys. Rev. Lett.*, 96:090601, 2006.

[17] M. Ceriotti, G. A. Tribello, and M. Parrinello. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci.*, 108(32):13023–13028, 2011.

[18] Eri Chatani, Rintaro Inoue, Hiroshi Imamura, Masaaki Sugiyama, Minoru Kato, Masahide Yamamoto, Koji Nishida, and Toshiji Kanaya. Early aggregation preceding the nucleation of insulin amyloid fibrils as monitored by small angle X-ray scattering. *Sci. Rep.*, 5:15485, oct 2015.

[19] Eliodoro Chiavazzo, Roberto Covino, Ronald R. Coifman, C. William Gear, Anastasia S. Georgiou, Gerhard Hummer, and Ioannis G. Kevrekidis. Intrinsic map dynamics exploration for uncharted effective free-energy landscapes. *Proc. Natl. Acad. Sci.*, page 201621481, 2017.

[20] Chi Cheng Chiu and Juan J. de Pablo. Fibrillar dimer formation of islet amyloid polypeptides. *AIP Adv.*, 5(9), 2015.

[21] Mikkel Christensen, Katrine K. Skeby, and Birgit Schitt. Identification of key interactions in the initial self-assembly of amylin in a membrane environment. *Biochemistry*, 56(36):4884–4894, 2017.

[22] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Multiscale methods. *Proc. Natl. Acad. Sci.*, 102(21):7432–7437, 2005.

[23] Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21(1):5–30, 2006.

[24] D. Cordell, A. Rosemarin, J.J. Schrder, and A.L. Smit. Towards global phosphorus security: A systems framework for phosphorus recovery and reuse options. *Chemosphere*, 84(6):747 – 758, 2011.

[25] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10092, 1993.

[26] Eric Darve, David Rodrguez-Gmez, and Andrew Pohorille. Adaptive biasing force method for scalar and vector free energy calculations. *The Journal of Chemical Physics*, 128(14):144120, 2008.

[27] Xavier Daura, Karl Gademann, Bernhard Jaun, Dieter Seebach, Wilfred F. van Gunsteren, and Alan E. Mark. Peptide folding: When simulation meets experiment. *Angewandte Chemie International Edition*, 38(12):236–240, 1999.

[28] Curt A. Davey, David F. Sargent, Karolin Luger, Armin W. Maeder, and Timothy J. Richmond. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J. Mol. Biol.*, 319(5):1097–1113, 2002.

[29] Gregory L. Dignon, Gl H. Zerze, and Jeetain Mittal. Interplay between membrane composition and structural stability of membrane-bound hiapp. *The Journal of Physical Chemistry B*, 121(37):8661–8668, 2017.

[30] Mojie Duan, Jue Fan, Minghai Li, Li Han, and Shuanghong Huo. Evaluation of dimensionality-reduction methods from peptide folding-unfolding simulations. *J. Chem. Theory Comput.*, 9(5):2490–2497, 2013.

[31] Ulrich Essmann, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. A smooth particle mesh Ewald method. *J. Chem. Phys.*, 103(19):8577–8593, 1995.

[32] Aaron M. Fluitt and Juan J. de Pablo. An Analysis of Biomolecular Force Fields for Simulations of Polyglutamine in Solution. *Biophys. J.*, 109(5):1009–1016, 2015.

[33] Gordon S. Freeman, Daniel M. Hinckley, Joshua P. Lequieu, Jonathan K. Whitmer, and Juan J. De Pablo. Coarse-grained modeling of DNA curvature. *J. Chem. Phys.*, 141(16), 2014.

[34] Joel M. Gottesfeld, Jason M. Belitsky, Christian Melander, Peter B. Dervan, and Karolin Luger. Blocking transcription through a nucleosome with synthetic DNA ligands. *J. Mol. Biol.*, 321(2):249–263, 2002.

[35] Stanley B. Grant, Jean-Daniel Saphores, David L. Feldman, Andrew J. Hamilton, Tim D. Fletcher, Perran L. M. Cook, Michael Stewardson, Brett F. Sanders, Lisa A. Levin, Richard F. Ambrose, Ana Deletic, Rebekah Brown, Sunny C. Jiang, Diego Rosso, William J. Cooper, and Ivan Marusic. Taking the "waste" out of "wastewater" for human water security and ecosystem sustainability. *Science*, 337(6095):681–686, 2012.

[36] Mathias Gruber, Per Greisen, Caroline M. Junker, and Claus Hlix-Nielsen. Phosphorus binding sites in proteins: Structural preorganization and coordination. *The Journal of Physical Chemistry B*, 118(5):1207–1215, 2014.

[37] Jeremy S. Guest, Steven J. Skerlos, James L. Barnard, M. Bruce Beck, Glen T. Daigger, Helene Hilger, Steven J. Jackson, Karen Karvazy, Linda Kelly, Linda Macpherson, James R. Mihelcic, Amit Pramanik, Lutgarde Raskin, Mark C. M. Van Loosdrecht, Daniel Yeh, and Nancy G. Love. A new planning and design paradigm to achieve sustainable resource recovery from wastewater. *Environmental Science & Technology*, 43(16):6126–6130, 2009.

[38] Ashley Z. Guo, Aaron M. Fluitt, and Juan J. de Pablo. Early-stage human islet amyloid polypeptide aggregation: Mechanisms behind dimer formation. *The Journal of Chemical Physics*, 149(2):025101, 2018.

[39] Ashley Z. Guo, Joshua Lequieu, and Juan J. de Pablo. Extracting collective motions underlying nucleosome dynamics via nonlinear manifold learning. *The Journal of Chemical Physics*, 150(5):054902, 2019.

[40] Ashley Z. Guo, Emre Sevgen, Hythem Sidky, Jonathan K. Whitmer, Jeffrey A. Hubbell, and Juan J. de Pablo. Adaptive enhanced sampling by force-biasing using neural networks. *J. Chem. Phys.*, 148(13):134108, 2018.

[41] James A. Hebda and Andrew D. Miranker. The interplay of catalysis and toxicity by amyloid intermediates on lipid bilayers: insights from type II diabetes. *Annu. Rev. Biophys.*, 38:125–52, 2009.

[42] B Hendrich and W Bickmore. Human diseases with underlying defects in chromatin structure and modification. *Hum. Mol. Genet.*, 10(20):2233–42, 2001.

[43] Berk Hess, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18(12):1463–1472, 1997.

[44] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. GROMACS 4 : Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.*, 4(3):435–447, 2008.

[45] Daniel M. Hinckley, Gordon S. Freeman, Jonathan K. Whitmer, and Juan J. De Pablo. An experimentally-informed coarse-grained 3-site-per-nucleotide model of DNA: Structure, thermodynamics, and dynamics of hybridization. *J. Chem. Phys.*, 139(14), 2013.

[46] Kyle Quynn Hoffmann, Michael Mcgovern, Chi-cheng Chiu, and Juan J de Pablo. Secondary Structure of Rat and Human Amylin across Force Fields. *PLoS One*, pages 1–24, 2015.

[47] Rundong Hu, Baiping Ren, Mingzhen Zhang, Hong Chen, Yonglan Liu, Lingyun Liu, Xiong Gong, Binbo Jiang, Jie Ma, and Jie Zheng. Seed-induced heterogeneous cross-seeding self-assembly of human and rat islet polypeptides. *ACS Omega*, 2(3):784–792, 2017.

[48] Rundong Hu, Mingzhen Zhang, Hong Chen, Binbo Jiang, and Jie Zheng. Cross-seeding interaction between β-amyloid and human islet amyloid polypeptide. *ACS Chemical Neuroscience*, 6(10):1759–1768, 2015.

[49] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: Visual Molecular Dynamics. *J. Mol. Graph.*, 14:33–38, 1996.

[50] Gaetano Invernizzi, Elena Papaleo, Raimon Sabate, and Salvador Ventura. Protein aggregation: Mechanisms and functional consequences. *Int. J. Biochem. Cell Biol.*, 44(9):1541–1554, 2012.

[51] I.T. Jolliffe. *Principal Component Analysis.* Springer Verlag, 1986.

[52] 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.

[53] Wolfgang Kabsch and Christian Sander. Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*, 22:2577–2637, 1983.

[54] Sang Beom Kim, Carmeline J. Dsilva, Ioannis G. Kevrekidis, and Pablo G. Debenedetti. Systematic characterization of protein folding pathways using diffusion maps: Application to Trp-cage miniprotein. *J. Chem. Phys.*, 142(8):085101, 2015.

[55] Thomas A. Knotts, Nitin Rathore, David C. Schwartz, and Juan J. De Pablo. A coarse grain model for DNA. *J. Chem. Phys.*, 126(8), 2007.

[56] I. M. Kulić and H. Schiessel. Chromatin dynamics: Nucleosomes go mobile through twist defects. *Phys. Rev. Lett.*, 91(14):3–6, 2003.

[57] Igor M. Kulić and H. Schiessel. Nucleosome repositioning via loop formation. *Biophys. J.*, 84(5):3197–3211, 2003.

[58] Shankar Kumar, Djamal Bouzida, Robert H. Swendsen, Peter A. Kollman, and John M. Rosenberg. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.*, 13(8):1011–1021, 1992.

[59] Alessandro Laio and Michele Parrinello. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.*, 99:12562, 2002.

[60] Joshua Lequieu, David C. Schwartz, and Juan J. de Pablo. In silico evidence for sequence-dependent nucleosome sliding. *Proceedings of the National Academy of Sciences*, 114(44):E9197–E9205, 2017.

[61] Gu Li, Marcia Levitus, Carlos Bustamante, and Jonathan Widom. Rapid spontaneous accessibility of nucleosomal DNA. *Nat. Struct. Mol. Biol.*, 12(1):46–53, 2005.

[62] W. Li, P. G. Wolynes, and S. Takada. Frustration, specific sequence dependence, and nonlinearity in large-amplitude fluctuations of allosteric proteins. *Proc. Natl. Acad. Sci.*, 108(9):3504–3509, 2011.

[63] Brian F. Lin, Katie A. Megley, Nickesh Viswanathan, Daniel V. Krogstad, Laurie B. Drews, Matthew J. Kade, Yichun Qian, and Matthew V. Tirrell. ph-responsive branched peptide amphiphile hydrogel designed for applications in regenerative medicine with potential as injectable tissue scaffolds. *J. Mater. Chem.*, 22:19447–19454, 2012.

[64] Kresten Lindorff-Larsen, Paul Maragakis, Stefano Piana, Michael P. Eastwood, Ron O. Dror, and David E. Shaw. Systematic validation of protein force fields against experimental data. *PLoS One*, 7(2):1–6, 2012.

[65] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L. Klepeis, Ron O. Dror, and David E. Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct. Funct. Bioinforma.*, 78(8):1950–1958, 2010.

[66] Na Liu, Mojie Duan, and Minghui Yang. Structural Properties of Human IAPP Dimer in Membrane Environment Studied by All-Atom Molecular Dynamics Simulations. *Sci. Rep.*, 7(1):7915, 2017.

[67] Andrew W. Long and Andrew L. Ferguson. Nonlinear machine learning of patchy colloid self-assembly pathways and mechanisms. *J. Phys. Chem. B*, 118(15):4228–4244, 2014.

[68] Y. Lorch, B. Davis, and R. D. Kornberg. Chromatin remodeling by DNA bending, not twisting. *Proc. Natl. Acad. Sci.*, 102(5):1329–1332, 2005.

[69] A. Lorenzo, B. Razzaboni, G C Weir, and B A Yankner. Pancreatic islet cell toxicity of amylin associated with type-2 diabetes mellitus. *Nature*, 368(6473):756–760, 1994.

[70] Sorin Luca, Wai Ming Yau, Richard Leapman, and Robert Tycko. Peptide conformation and supramolecular organization in amylin fibrils: Constraints from solid-state NMR. *Biochemistry*, 46(47):13505–13522, 2007.

[71] Karolin Luger, Armin W. Mäder, Robin K. Richmond, David F. Sargent, and Timothy J. Richmond. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648):251–260, 1997.

[72] Thomas A. Lutz. Control of energy homeostasis by amylin. *Cell. Mol. Life Sci.*, 69(12):1947–1965, 2012.

[73] M. Maj, J.P. Lomont, K.L. Rich, A.M. Alperstein, and M.T. Zanni. Site-specific detection of protein secondary structure using 2D IR dihedral indexing: A proposed assembly mechanism of oligomeric hIAPP. *Chem. Sci.*, 9(2):463–474, 2018.

[74] Luca Maragliano, Eric Vanden-Eijnden, and Benoît Roux. Free energy and kinetics of conformational transitions from voronoi tessellated milestoning with restraining potentials. *J. Chem. Theory Comput.*, 5(10):2589–2594, 2009.

[75] Anne Martel, Lucas Antony, Yuri Gerelli, Lionel Porcar, Aaron Fluitt, Kyle Hoffmann, Irena Kiesel, Michel Vivaudou, Giovanna Fragneto, and Juan J. de Pablo. Membrane permeation versus amyloidogenicity: A multitechnique study of islet amyloid polypeptide interaction with model membranes. *Journal of the American Chemical Society*, 139(1):137–148, 2017.

[76] Chirag M. Mehta, Wendell O. Khunjar, Vivi Nguyen, Stephan Tait, and Damien J. Batstone. Technologies to recover nutrients from waste streams: A critical review. *Critical Reviews in Environmental Science and Technology*, 45(4):385–427, 2015.

[77] E. James Milner-White and Michael J. Russell. Sites for phosphates and iron-sulfur thiolates in the first membranes: 3 to 6 residue anion-binding motifs (nests). *Origins of Life and Evolution of Biospheres*, 35(1):19–27, Feb 2005.

[78] Tajib A. Mirzabekov, Meng Chin Lin, and Bruce L. Kagan. Pore formation by the cytotoxic islet amyloid peptide amylin. *J. Biol. Chem.*, 271(4):1988–1992, 1996.

[79] Daniel F. Moriarty and Daniel P. Raleigh. Effects of sequential proline substitutions on amyloid formation by human amylin20-29. *Biochemistry*, 38(6):1811–1818, 1999.

[80] Gregg B. Morin. The human telomere terminal transferase enzyme is a ribonucleoprotein that synthesizes TTAGGG repeats. *Cell*, 59(3):521–529, 1989.

[81] R. K. Moyzis, J. M. Buckingham, L. S. Cram, M. Dani, L. L. Deaven, M. D. Jones, J. Meyne, R. L. Ratliff, and J. R. Wu. A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. *Proc. Natl. Acad. Sci.*, 85(18):6622–6626, 1988.

[82] Larissa A Munishkina and Anthony L Fink. Fluorescence as a method to reveal structures and membrane-interactions of amyloidogenic proteins. *Biochim. Biophys. Acta*, 1768:1862–1885, 2007.

[83] Praveen Nedumpully-Govindan, Esteban N. Gurzov, Pengyu Chen, Emily H. Pilkington, William J. Stanley, Sara A. Litwak, Thomas P. Davis, Pu Chun Ke, and Feng Ding. Graphene oxide inhibits hiapp amyloid fibrillation and toxicity in insulin-producing nit-1 cells. *Phys. Chem. Chem. Phys.*, 18:94–100, 2016.

[84] Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.*, 81(1):511–519, 1984.

[85] Alexey Onufriev, Donald Bashford, and David A. Case. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Structure, Function, and Bioinformatics*, 55(2):383–394, 2004.

[86] Chris Oostenbrink, Alessandra Villa, Alan E. Mark, and Wilfred F. Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.*, 25(13):1656–1676, 2004.

[87] Marco Pasi and Richard Lavery. Structure and dynamics of DNA loops on nucleosomes studied with atomistic, microsecond-scale molecular dynamics. *Nucleic Acids Res.*, 44(11):5450–5456, 2016.

[88] S. Piana, K. Lindorff-Larsen, and D. E. Shaw. Protein folding kinetics and thermodynamics from atomistic simulation. *Proc. Natl. Acad. Sci.*, 109(44):17845–17850, 2012.

[89] Stefano Piana and Alessandro Laio. A Bias-Exchange Approach to Protein Folding. *J. Phys. Chem. B*, 111(17):4553–4559, 2007.

[90] Fabio Pietrucci and Alessandro Laio. A Collective Variable for the Efficient Exploration of Protein Beta-Sheet Structures : Application to SH3 and GB1. *J. Chem. Theory Comput.*, 5(9):2197–2201, 2009.

[91] Yair Porat, Sofiya Kolusheva, Raz Jelinek, and Ehud Gazit. The human islet amyloid polypeptide forms transient membrane-active prefibrillar assemblies. *Biochemistry*, 42(37):10971–10977, 2003.

[92] Yair Porat, Yariv Mazor, Shimon Efrat, and Ehud Gazit. Inhibition of islet amyloid polypeptide fibril formation: A potential role for heteroaromatic interactions. *Biochemistry*, 43(45):14454–14462, 2004.

[93] P. Ranjith, J. Yan, and J. F. Marko. Nucleosome hopping and sliding kinetics determined from dynamics of single chromatin fibers in Xenopus egg extracts. *Proc. Natl. Acad. Sci.*, 104(34):13649–13654, 2007.

[94] T J Richmond and C A Davey. The structure of DNA in the nucleosome core. *Nature*, 423(6936):145–150, 2003.

[95] Bruce E. Rittmann, Brooke Mayer, Paul Westerhoff, and Mark Edwards. Capturing the lost phosphorus. *Chemosphere*, 84(6):846 – 853, 2011.

[96] Mary A. Rohrdanz, Wenwei Zheng, Mauro Maggioni, and Cecilia Clementi. Determination of reaction coordinates via locally scaled diffusion map. *J. Chem. Phys.*, 134(12), 2011.

[97] Sam Roweis and L Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(December):2323–2326, 2000.

[98] E. J. Sambriski, D. C. Schwartz, and J. J. De Pablo. A mesoscale model of DNA and its renaturation. *Biophys. J.*, 96(5):1675–1690, 2009.

[99] Matti Saraste, Peter R. Sibbald, and Alfred Wittinghofer. The p-loop a common motif in atp- and gtp-binding proteins. *Trends in Biochemical Sciences*, 15(11):430 – 434, 1990.

[100] Silvia Scalisi, Michele F M Sciacca, Genady Zhavnerko, Domenico M. Grasso, Giovanni Marletta, and Carmelo La Rosa. Self-assembling pathway of HiApp fibrils within lipid bilayers. *ChemBioChem*, 11(13):1856–1859, 2010.

[101] H. Schiessel, J. Widom, R. F. Bruinsma, and W. M. Gelbart. Polymer reptation and nucleosome repositioning. *Phys. Rev. Lett.*, 86(19):4414–4417, 2001.

[102] Eran Segal, Yvonne Fondufe-Mittendorf, Lingyi Chen, Annchristine Thåström, Yair Field, Irene K. Moore, Ji Ping Z Wang, and Jonathan Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778, 2006.

[103] Dennis J Selkoe. Folding proteins in fatal ways. *Nature*, 426(6968):900–904, 2003.

[104] Arnaldo L. Serrano, Justin P. Lomont, Ling-Hsien Tu, Daniel P. Raleigh, and Martin T. Zanni. A free energy barrier caused by the refolding of an oligomeric intermediate controls the lag time of amyloid formation by hiapp. *Journal of the American Chemical Society*, 139(46):16748–16758, 2017.

[105] Emre Sevgen, Ashley Guo, Hythem Sidky, Jonathan K Whitmer, and Juan J. de Pablo. Combined force-frequency sampling for simulation of systems having rugged free energy landscapes. *Journal of Chemical Theory and Computation*, 2020.

[106] Thomas E Shrader and Donald M Crothers. Artificial nucleosome positioning sequences (chromatin/histone-DNA binding/DNA bending). *Biophysics (Oxf).*, 86(October):7418–7422, 1989.

[107] H. Sidky, Y.J. Colón, J. Helfferich, B.J. Sikora, C. Bezik, W. Chu, F. Giberti, A.Z. Guo, X. Jiang, J. Lequieu, J. Li, J. Moller, M.J. Quevillon, M. Rahimi, H. Ramezani-Dakhel, V.S. Rathee, D.R. Reid, E. Sevgen, V. Thapar, M.A. Webb, J.K. Whitmer, and J.J. De Pablo. SSAGES: Software Suite for Advanced General Ensemble Simulations. *J. Chem. Phys.*, 148(4), 2018.

[108] Hythem Sidky, Yamil J. Coln, Julian Helfferich, Benjamin J. Sikora, Cody Bezik, Weiwei Chu, Federico Giberti, Ashley Z. Guo, Xikai Jiang, Joshua Lequieu, Jiyuan Li, Joshua Moller, Michael J. Quevillon, Mohammad Rahimi, Hadi Ramezani-Dakhel, Vikramjit S. Rathee, Daniel R. Reid, Emre Sevgen, Vikram Thapar, Michael A. Webb, Jonathan K. Whitmer, and Juan J. de Pablo. Ssages: Software suite for advanced general ensemble simulations. *The Journal of Chemical Physics*, 148(4):044104, 2018.

[109] Ralf Strohner, Malte Wachsmuth, Karoline Dachauer, Jacek Mazurkiewicz, Julia Hochstatter, Karsten Rippe, and Gernot Längst. A 'loop recapture' mechanism for ACF-dependent nucleosome remodeling. *Nat. Struct. Mol. Biol.*, 12(8):683–690, 2005.

[110] Robert K. Suto, Rajeswari S. Edayathumangalam, Cindy L. White, Christian Melander, Joel M. Gottesfeld, Peter B. Dervan, and Karolin Luger. Crystal structures of nucleosome core particles in complex with minor groove DNA-binding ligands. *J. Mol. Biol.*, 326(2):371–380, 2003.

[111] Shoji Takada, Zaida Luthey-Schulten, and Peter G. Wolynes. Folding dynamics with nonadditive forces: A simulation study of a designed helical protein and a random heteropolymer. *J. Chem. Phys.*, 110(23):11616–11629, 1999.

[112] J B Tenenbaum, V de Silva, and J C Langford. A global geometric framework for non linear dimensionality reduction. *Science*, 290(December):2319–2323, 2000.

[113] Hannah S. Tims, Kaushik Gurunathan, Marcia Levitus, and Jonathan Widom. Dynamics of nucleosome invasion by DNA binding proteins. *J. Mol. Biol.*, 411(2):430–448, 2011.

[114] Amanda Trent, Rachel Marullo, Brian Lin, Matthew Black, and Matthew Tirrell. Structural properties of soluble peptide amphiphile micelles. *Soft Matter*, 7:9572–9582, 2011.

[115] Eric Vanden-Eijnden and Maddalena Venturoli. Revisiting the finite temperature string method for the calculation of reaction tubes and free energies. *J. Chem. Phys.*, 130(19):1–17, 2009.

[116] Jiang Wang, Mohit A. Gayatri, and Andrew L. Ferguson. Mesoscale Simulation and Machine Learning of Asphaltene Aggregation Phase Behavior and Molecular Assembly Landscapes. *J. Phys. Chem. B*, 121(18):4923–4944, 2017.

[117] Lu Wang, Chris T. Middleton, Sadanand Singh, Allam S. Reddy, Ann M. Woys, David B. Strasfeld, Peter Marek, Daniel P. Raleigh, Juan J. de Pablo, Martin T. Zanni, and James L. Skinner. 2DIR spectroscopy of human amylin fibrils reflects stable β-sheet structure. *J. Am. Chem. Soc.*, 133(40):16062–16071, 2011.

[118] James D Watson and E.James Milner-White. A novel main-chain anion-binding site in proteins: the nest. a particular combination of , values in successive residues gives rise to anion-binding sites that occur commonly and are found often at functionally important regions11edited by j. thornton. *Journal of Molecular Biology*, 315(2):171 – 182, 2002.

[119] P Westermark, A Andersson, and G T Westermark. Islet amyloid polypeptide, islet amyloid, and diabetes mellitus. *Physiol Rev*, 91(3):795–826, 2011.

[120] Per Westermark, U Engström, K Johnson, Gunilla Westermark, and Christer Betsholtz. Islet amyloid polypeptide: pinpointing amino acid residues linked to amyloid fibril formation. *Proc. Natl. Acad. Sci. U. S. A.*, 87(July):5036–40, 1990.

[121] Vered Wineman-Fisher and Yifat Miller. Insight into a new binding site of zinc ions in fibrillar amylin. *ACS Chemical Neuroscience*, 8(9):2078–2087, 2017.

[122] Mingzhen Zhang, Baiping Ren, Hong Chen, Yan Sun, Jie Ma, Binbo Jiang, and Jie Zheng. Molecular simulations of amyloid structures, toxicity, and inhibition. *Israel Journal of Chemistry*, 57(7-8):586–601, 2017.

[123] Mingzhen Zhang, Baiping Ren, Yonglan Liu, Guizhao Liang, Yan Sun, Lijian Xu, and Jie Zheng. Membrane interactions of hiapp monomer and oligomer with lipid membranes by molecular dynamics simulations. *ACS Chemical Neuroscience*, 8(8):1789–1800, 2017.

[124] Wenwei Zheng, Mary A. Rohrdanz, and Cecilia Clementi. Rapid exploration of configuration space with diffusion-map-directed molecular dynamics. *J. Phys. Chem. B*, 117(42):12769–12776, 2013.