Mechanistic insights into proteins association through molecular dynamics

Chua, Khi Pin

2017

Chua, K. P. (2017). Mechanistic insights into proteins association through molecular dynamics. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/73025>

<https://doi.org/10.32657/10356/73025>

NANYANG
TECHNOLOGICAL
UNIVERSITY

**Mechanistic Insights Into Proteins Association
Through Molecular Dynamics**

Khi Pin Chua
Interdisciplinary Graduate School
NTU Institute for Health Technologies
2017

Mechanistic Insights Into Proteins Association Through Molecular Dynamics

Khi Pin Chua

Interdisciplinary Graduate School
NTU Institute for Health Technologies

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of
Doctor of Philosophy

2017

Acknowledgments

I would like to express the utmost gratitude to my supervisor, Associate Professor Chew Lock Yue for his guidance throughout the four years of my doctoral study. He has been the biggest source of motivation for me. His insistence in doing science the right way is an inspiration to all of us in the lab. Additionally, I would like to thank my co-supervisor, Associate Professor Mu Yuguang for steering me in the right direction when I was lost. His experience in the molecular dynamics field has been one of the main factors I learnt so much during the past four years.

Other than that, I would like to thank my mentor, Associate Professor Konstantin Pervushin for giving me valuable feedbacks in the projects I have done during our meetings. The feedbacks were crucial in polishing my works. My sincere thanks also go to my collaborators Assistant Professor Miao Yansong, Associate Professor Ali Miserez, Sun He and Hiew Shu Hui. I have learnt from them the nuances of experimental biology during our interdisciplinary collaborations.

I would also like to take this chance to thank my labmates for the useful discussions during group meetings and for all the fun we had going out for group dinner. I could not have gone through the thesis without their supports.

I am extremely grateful to my family who has supported me unconditionally throughout the thesis. They have provided me moral and emotional support when I had doubt in this journey. Lastly, I would like to thank my friends for the social life I needed besides working in the lab. And the friend who had the patience to listen to a nerdy scientists' random science facts and his struggles with life.

Contents

Acronyms	iv
List of Figures	v
List of Tables	ix
1 Overview and Main Methodology	1
1.1 Basic Structures of Proteins and Protein Folding Problem	1
1.2 Proteins' Aggregation and Gelation	5
1.3 Protein-protein Recognition	7
2 Methodology and Literatures Review	9
2.1 Molecular Dynamics	9
2.2 Applications of Molecular Dynamics on Protein-protein Interactions . .	14
2.2.1 Amylin (IAPP) Aggregation	14
2.2.2 Prion Fragment: PrP106-126 Aggregation	16
2.2.3 Peptides Gelation	19
2.2.4 Protein-protein Complexes: Recognition of Polyprolines	22
2.3 Aims and Objectives	24
3 Replica Exchange Molecular Dynamics Simulation of Cross-Fibrillation of IAPP and PrP106-126	26
3.1 Introduction	26
3.2 Simulation Details and Analysis Methods	28
3.3 Results	31
3.3.1 REMD Diagnostic Tests	31
3.3.2 Secondary Structures Propensity and Contact's Probability . . .	32
3.3.3 Dihedral Principal Component Analysis	35
3.3.4 β -sheets Formation	40
3.4 Discussion and Conclusion	43
3.5 Supplementary Information	47
3.5.1 Comparison of dPCA Energy Landscape at Three Lowest Temperatures	47
3.5.2 The Evolution of into Structure E	47
3.5.3 DSSP Plot for 315 K Ensemble	50
3.5.4 Dynamical Reweighting using MBAR	50
3.5.5 Structures from Conformational Clustering	53
3.5.6 Constant Temperature Molecular Dynamics with Amber99SB-ILDN	55
4 Towards a Mechanistic Understanding of Peptides Gelation using Sequences from Sucker Ring Teeth Proteins	56
4.1 Introduction	56
4.2 Methods	58
4.2.1 Martini Coarse-grained Model	58

4.2.2	Bias Exchange Well-tempered Metadynamics (WT-BEMD) . . .	59
4.3	Results	62
4.3.1	Coarse-grained Simulations using MARTINI Model	62
4.3.2	Convergence of Metadynamics Simulation	67
4.3.3	Metadynamics: Difference in Free Energy Landscape	70
4.3.4	Metadynamics: Multi-dimensional Free Energy Clustering Analysis	71
4.4	Discussion	76
4.5	Supplementary Figures	79
5	Binding Analysis of Profilin-Polyproline Complex	81
5.1	Introduction	81
5.2	Methods	84
5.2.1	Homology Modelling and Molecular Dynamics	84
5.2.2	MM/PBSA Binding Energy Analysis	86
5.3	Results	88
5.4	Discussions	93
5.4.1	Stability of AtPRF3-Polyprolines Complexes and Extended N-terminal α -helix	93
5.4.2	CH- π and Electrostatic Interactions	94
5.4.3	MM/PBSA Binding Free Energy	96
5.5	Supplementary Information	98
6	Conclusion and Future Outlook	102
7	List of Publications	107
8	References	108

Acronyms

A β Amyloid Beta protein.

AtPRF *Arabidopsis Thaliana* Profilin.

BEMD Bias-exchange Metadynamics.

CD Circular Dichroism.

dPCA Dihedral Principal Component Analysis.

FH1 Formin Homology Domain 1.

hIAPP Human Islet Amyloid Polypeptide.

IDP Intrinsically disordered protein.

MD Molecular Dynamics.

METAD Metadynamics.

MM/PBSA Molecular Mechanics/Poisson Boltzmann Surface Area.

NMR Nuclear Magnetic Resonance.

PDB Protein Data Bank.

PMF Potential mean force.

PolyP Poly-prolines.

PP-II Poly-prolines Type II.

PrP Prion Protein.

PrP106-126 Prion protein fragment (residue 106 to residue 126).

REMD Replica-exchange Molecular Dynamics.

RMSD Root-mean-square Deviation.

SASA Solvent-accessible Surface Area.

SH2/SH3 Src Homology Domain 2/3.

SRT Sucker Ring Teeth.

VdW Van der Waal.

WT-BEMD Well-tempered Bias-exchange Metadynamics.

List of Figures

1	Illustrations of a/an (a) α -helix. (b) Anti-parallel β -sheets. (c) Parallel β -sheets. Figure (a) was generated by author using PyMOL. ¹ Figures (b) and (c) were obtained from the Internet from Wikimedia Common Repository. ^{2,3}	3
2	Figure shows a protofibril of amyloid beta ($A\beta$) solved experimentally. Generated using PyMOL using structure 2BEG from Protein Data Bank (PDB). ^{1,4}	5
3	(a) Structure of SH3 with a proline-rich peptide (PDB code:1PRL). ⁵ (b) Structure of a hemoglobin heterotetramer (PDB code:1C7C). ⁶ Figures were generated by the author using PyMOL. ¹	8
4	(a) Crystal structure of full length human amylin (hIAPP). (b) NMR structure of anti-parallel β -sheets structures of hIAPP20-29 (SNNF-GAILS). Figures are generated by author using PyMOL. ^{1,7,8}	15
5	(a) Solution NMR structure of human prion protein (PDB Code:1QLX). ⁹ Structures shown are for residue 125 to residue 228. (b) X-ray crystal structure of dimeric prion. ¹⁰ Figure generated by author using PyMOL. ¹	18
6	Crystal structure of the N-terminal domain of silk fibroin from <i>Bombyx mori</i> . Figure generated by author using PyMOL. ^{1,11}	21
7	Test for validity of REMD simulation. (A) Sufficient overlaps are found between the potential energy distribution of all the temperatures used in the simulation. Note that each color represents the distribution from a single temperature. (B) Random walk in the temperature space for all replicas (y-axis). Colors indicate temperature ensembles, ranging from lowest temperature (blue) to highest temperature (red). (C) The average coil propensity at different temperatures are similar for ten periods during the simulation. (D) Average turn propensity which has been evaluated in a similar fashion as (C). (E) Evolution of secondary structures for the 315 K ensemble. (F) An estimate of the conformational entropy of the protein complex in bound state by the quasi-harmonic method.	32
8	Propensity of the different secondary structures adopted by PrP106-126 and IAPP during the 100-600 ns period. The gray bar indicated by CS separates the two chains. Index 106-126 is PrP106-126 while index 1-37 is IAPP.	33
9	Frequency of contacts formed by (A) PrP106-126 towards IAPP and (B) IAPP towards PrP106-126. (C) Mean distance between residues of IAPP and PrP106-126. All contacts and distances were calculated by including only the heavy atoms.	35
10	Contour plot of the relative free energy landscape determined from dPCA using the first and second principal components of REMD simulation at 315K. Note that the color bar represents the relative free energy (kJ mol^{-1}), which is evaluated from the population of conformations obtained by determining the frequency of the conformation within a grid of the energy landscape.	36

11	(A-F) Representative structures for six minima in the dPCA relative free energy landscape. Note that red color is used for IAPP, while blue color is used for PrP106-126. Yellow spheres indicate the N-termini. Figures were generated with PyMOL. (G-H) Representative structure from the 9th cluster determined from conformational clustering. (I) Secondary structures determined by DSSP for A-G. Z is the initial configuration of the system. Yellow indicates β -sheets, gold indicates β -bridge, orange indicates 3-10 helix and cyan indicates α -helix.	37
12	(1-8) Snapshots containing inter-chain β -sheets structures. Figures were generated using PyMOL. (9) Secondary structures determined by DSSP for structures 1-8. The definition of colors and representation of structures are identical to those given in Figure 11.	41
S1	Comparison of the dPCA energy landscape for the three lowest temperatures ensembles.	47
S2	Snapshots of the different stages described in the structural evolution of PrP106-126 and IAPP complex. Yellow ball indicates N-terminal. Blue is used for PrP106-126 chain, while red is used for IAPP chain. t denotes the time of the frame, T indicates the temperature bath the replica was in at the particular time. Figures were generated with PyMOL. ¹ . . .	49
S3	DSSP plot for ensemble 0 showing the diverse sampling throughout the REMD simulation.	50
S4	2D potential of mean force plot using dynamical reweighting on all 55 ensembles from 100 ns to 500 ns.	52
S5	Top 9 structures from conformation clustering method described in main manuscript. The numbering of the structures corresponds to the population number in descending order.	53
S6	Percentage population of all replicas visiting the middle (30th ensemble of the 55 total ensemble) ensemble simulated at 382 K. This shows that replicas from both left and right of the ensemble visited the middle ensemble with a random pattern in general.	54
S7	Root-mean-square deviation of distances between backbone atoms for structure A-F obtained from dPCA analysis and then simulated using Amber99SB-ILDN using constant temperature molecular dynamics. . .	55
13	Solvent Accessible Surface Area (SASA) for the different peptide systems simulated. Each color represents a repeat of the simulation for the same peptide system. (a) GV8. (b) GL8. (c) GV8 at 2000 mM. (d) GL8 at 2000 mM.	64
14	Radial distribution function $g(r)$ for (a) GV8 (b) GL8 (c) GV8 at 2000 mM and (d) GL8 at 2000 mM concentration.	65
15	Representative structures for (a)(c) Run one to run three of peptide GV8 in 400 mM concentration and (d)(f) Run one to three of peptide GL8 in 400 mM concentration. Water molecules are not shown here as they are not important to the visualization. (g) Scanning electron microscope (SEM) images of both peptides GV8 and GL8 (Image provided by Shu Hui, Hiew and Ali Miserez).	66

16	Representative structures for the final structures obtained in simulations with 2000 mM concentration of (a) peptides GV8 and (b) peptides GL8. Water molecules are colored in cyan. Periodic boundary conditions were shown by drawing the neighbouring boxes in x, y and z direction. . . .	67
17	Free energy for the 7 CVs used in the simulation for peptide GV8. Magenta line represents free energy sampled during the first 200 ns of the last 400 ns of the simulations, while blue line represents free energy sampled during the second 200 ns of the last 400 ns of the simulations. Green line represents the average value of the magenta and blue line, with the additional condition that it will only be drawn if the deviation between magenta and blue line is lower than $6 k_B T$ (a) AlphaRMSD. (b) Anti-parallel β -RMSD. (c) Parallel β -RMSD. (d) Side-chain contact of Leucine and Valine. (e) Backbone hydrogen bonds. (f) Contacts between α -carbon. (g) Radius of gyration.	69
18	Free energy for the 7 CVs used in the simulation for peptide GL8. Color of the lines are defined to be the same as that of figure 17. (a) AlphaRMSD. (b) Anti-parallel β -RMSD. (c) Parallel β -RMSD. (d) Side-chain contact of Leucine and Valine. (e) Backbone hydrogen bonds. (f) Contacts between α -carbon. (g) Radius of gyration.	70
19	β -sheets propensity for every amino acid in the sequence. Error bars were calculated as the standard deviation of average values from four monomer for each amino acid.	73
20	Figure shows the ensemble average of (a) β -sheets structures (b) number of hydrogen bonds at clusters below different free energy cutoffs. . . .	74
21	Representative structures from different energy minima. Arrows indicate the locations on the free energy landscape where the structures belong to.	75
S8	Figures show the potentials deposited throughout the 1.2 μs simulation for peptide GV8.	79
S9	Figures show the potentials deposited throughout the 1.2 μs simulation for peptide GL8.	80
22	Chemical composition of proline.	83
23	(a) Example of the initial binding positions of polyprolines to AtPRF3-N37. (b) Root means square deviation of distances between backbone atoms for homology model generated using SWISS-MODEL.	85
24	Figures on top show the representative structures for the largest cluster for simulation of AtPRF3-N37 in complex with polyprolines. Figures at the bottom show the corresponding contact maps between residues of AtPRF3-N37 (y-axis) and polyprolines (x-axis).	90
25	Figures on top show the representative structures for the largest cluster for simulation of AtPRF3-N22 in complex with polyprolines. Figures at the bottom shows the corresponding contact maps between residues of AtPRF3-N22 (y-axis) and polyprolines (x-axis).	91

26	Figures on top show the representative structures for the second and third largest cluster for simulation of AtPRF3-N22 in complex with polyprolines. Figures at the bottom shows the corresponding contact maps between residues of AtPRF3-N22 (y-axis) and polyprolines (x-axis).	92
S10	Prediction of secondary structures for AtPRF3-N22 using PSIPRED webserver.	98
S11	RMS Deviation of distance of the backbone atoms of the complexes during simulation for (a) N37L-CN (b) N37U-CN (c) N37L-NC (d) N37U-NC (e) N22L-CN (f) N22U-CN	99
S12	(a) Peptide dissociation in N22L-NC. Other than dissociating from the binding groove, N-termini can also be seen to have rotated to the opposite orientation compared to the initial position (figure 23). (b) Unfolded N-terminal helix in N22U-NC. Cyan color represents the N-terminal extension. Blue sphere indicates the N-termini of polyprolines. Figures are generated by author using PyMOL. ¹	100

List of Tables

1	Table of 20 types of amino acids categorised by the physical properties of side chains.	2
2	Assignment of index to residues. First row is the indexes for PrP106-126 while the second and third rows contain indexes for IAPP. Note that index 22 is reserved as a chain separator to ease visualization. . .	27
3	Binding free energy estimated using MM/PBSA approach. The energies are rounded to one decimal place and shown in kilocalories per mol. The binding energy is decomposed into individual component by g_mmpbsa where ΔE_{elec} is the molecular mechanics (MM) electrostatic energy, ΔE_{vdW} is the MM van der Wal's energy, ΔG_{polar} is the polar solvation energy and $\Delta G_{nonpolar}$ is the non-polar solvation energy.	38
4	Binding free energy estimated using MM/PBSA approach with the individual components shown. The energies are rounded to one decimal place and shown in kilocalories per mol. The binding energy is decomposed into individual component by g_mmpbsa where ΔE_{elec} is the molecular mechanics (MM) electrostatic energy, ΔE_{vdW} is the MM van der Wal's energy, ΔG_{polar} is the polar solvation energy and $\Delta G_{nonpolar}$ is the non-polar solvation energy.	42
S1	Binding free energy estimated using MM/PBSA approach. Data for structures a, b, c, d and h were taken from table 2. The binding energy was decomposed into individual component by g_mmpbsa where ΔE_{elec} is the molecular mechanics (MM) electrostatic energy, ΔE_{vdW} is the MM van der Wals energy, ΔG_{polar} is the polar solvation energy and $\Delta G_{nonpolar}$ is the non-polar solvation energy.	54
S2	Average hydrogen bond number between IAPP and PrP106-126 for structure simulated with OPLS-AA forcefield. The value was calculated using g_hbond utility. Note that structures A-F were obtained from our REMD simulation through dPCA analysis.	55
S3	Average hydrogen bond number between IAPP and PrP106-126 for structure simulated with Amber99SB-ILDN forcefield. The value was calculated using g_hbond utility. Note that structures A-F were obtained from our REMD simulation through dPCA analysis.	55
5	Table shows moments of inertia for the final snapshots of simulated peptides GV8 and GL8 at 400 mM concentration. Units for radii of gyration are in nm . Units for moments of inertia are in $a.m.u.nm^2$. . .	65
6	Ensemble average values for the secondary structures explored during the simulation. Number indicates the average number of residues adopting the specific secondary structure. Errors are omitted here for clarity sake as they are negligible.	73
7	Table shows the α -helix content of the constructed N-terminal extension QRRSRAKVKKKKKTN . Percentage content was calculated as $\frac{\text{Number of } \alpha\text{-helix}}{15}$ where 15 is the length of the extension. Errors were estimated from the standard deviation during the last 25 ns of each run.	94

S4	Binding free energy estimated using MM/PBSA for N37L-CN and N37U-CN configurations. Unit for energy is kJ/mol.	100
S5	Binding free energy estimated using MM/PBSA for N37L-NC and N37U-NC configurations. Unit for energy is kJ/mol.	100
S6	Binding free energy estimated using MM/PBSA for N22L-CN and N22U-CN configurations. Unit for energy is kJ/mol.	100
S7	Binding free energy estimated using MM/PBSA for N22L-NC and N22U-NC configurations. Unit for energy is kJ/mol.	101
S8	Full FASTA sequence of formin FH1 domain of <i>arabidopsis thaliana</i> . .	101

List of Equations

2.1	Newton's equation of motion	12
2.2	Potential energy and force of molecular dynamics system	12
2.3	Force of molecular dynamics system	12
2.4	Set of energy terms describing molecular dynamics system.	12
2.5	Verlet integration method	13
3.1	Exchange probability of replica exchange molecular dynamics simulations.	28
4.2	Equation for metadynamics	60
4.3	Equation describing the collective variable (CV) used to quantify secondary structures in metadynamics	60
4.4	Radius of gyration	61
4.5	Moment of inertia	63
4.7	Ensemble average values of a metadynamics simulation and the corre- sponding errors.	72

Mechanistic Insights Into Proteins Association Through Molecular Dynamics

by

Khi Pin Chua

Abstract

The study of proteins' structures and dynamics is critical to understanding the question of "What is Life?". Human body, for example, contains innumerable proteins working in tandem to perform biological functions necessary for human survival. This thesis attempts to shed insights into various types of association and aggregation processes between proteins and peptides. In the first part of this thesis, we first look into the aggregation process of human islet amyloid polypeptide (hIAPP) and a small fragment of prion protein (106-126). These are two structurally and sequentially distinct peptides that were found to aggregate in experiments (Gal et al. JACS 2013, 135, 13582-9).¹² We discovered that the two peptides adopt highly polymorphic complex and can potentially form β -sheet rich structures. The second part of this thesis looks into the binding process of a profilin sequence from Arabidopsis Thaliana (AtPRF3) with a short polyproline peptide. Our binding energy analysis was consistent with that of experiments (Manuscript in preparation) and we were able to provide insights into the binding mechanism in atomistic details. Several important interactions were found such as the CH- π stacking and electrostatic attraction/repulsion. Lastly, this thesis concludes with a project investigating two short peptides from sucker ring teeth (SRT) proteins that differ only by a single amino acid. One of the peptides was able to form hydrogel in water while the other could not. By using large scale simulation together with coarse-grained modelling, we discovered fibril-like structures containing β -sheets.

1 Overview and Main Methodology

1.1 Basic Structures of Proteins and Protein Folding Problem

Proteins are the second most abundant molecules in the human body, accounting for $\approx 20\%$ of the mass. Proteins carry out various roles vital to human survival such as acting as enzymes, hormones, and immunoproteins etc. On that account, protein science is understandably one of the most well-studied fields as scientists seek to understand the mechanisms and dynamics of the countless proteins relevant to the physiological functions of living organisms. Despite the enormous efforts put into decades of studying proteins, many aspects remain a mystery to scientists due to the sheer complexity of proteins. Chief among these is the protein folding problem, which is a central theme of this thesis.

Traditionally, the structure of proteins is classified into four levels: primary, secondary, tertiary, and quaternary structures. The basic building block of the primary structure is called amino acids. There are 20 types of naturally occurring amino acids.¹³ Chemically, an amino acid consists of an alpha carbon (C_α) attached to an amine (NH_2) group and a carboxylic ($COOH$) group. A side chain group is attached to the alpha carbon. Naturally, the 20 types of amino acids are different from each other based on the chemical constituent of their side chain group. They are commonly categorised into polar, non-polar and charged amino acids (Table 1). The primary structure of a protein is defined as the linear sequence of amino acids (commonly referred to residues) forming the polypeptide chain. Note that in this thesis we will use peptide and polypeptide interchangeably to refer to a chain of two or more amino acids linked together via peptide bonds. Traditionally, the nomenclature “peptide” refers to short chain of amino acids while “polypeptide” refers to longer chain. The distinction however is unclear and deemed unimportant. In addition, a protein is defined as a folded peptide with well-defined secondary and tertiary structures (explained below).

Table 1. Table of 20 types of amino acids categorised by the physical properties of side chains.

Hydrophobic	Polar	Positively-charged	Negatively-charged
Glycine, GLY, G Alanine, ALA, A Valine, VAL, V Leucine, LEU, L Isoleucine, ILE, I Methionine, MET, M Phenylalanine, PHE, F Tryptophan, TRP, W Proline, PRO, P	Serine, SER, S Threonine, THR, T Cysteine, CYS, C Tyrosine, TYR, Y Asparagine, ASN, N Glutamine, GLN, Q	Lysine, LYS, K Arginine, ARG, R Histidine, HIS, H	Aspartic Acid, ASP, D Glutamic Acid, GLU, E

Next, the secondary structures of proteins refer to the local structures adopted by the polypeptide chain. These local structures are formed by backbones of a proteins. The most commonly known secondary structures are α -helix and β -sheets.¹⁴ Secondary structures are defined based on patterns of hydrogen bonds formed between the backbone atoms of proteins and peptides. An α -helix is defined as a contiguous segment of one peptide where the backbone oxygen and hydrogen atom of the i^{th} residue form hydrogen bonds with those of the $(i + 4)^{th}$ residue. On the other hand, β -sheet is defined as at least two hydrogen bonds formed between at least two strands of peptides. Unlike an α -helix, a β -sheets can be formed between two separate chains of peptides. Secondary structures are illustrated in Figure 1. β -sheets will be a recurring theme throughout this thesis as it plays a crucial role in stabilising protein-protein interactions. Thereafter, the arrangement of the various secondary structures of a polypeptide chain is defined as the tertiary structures. Lastly, multiple protein subunits can form a large functional unit referred to as the quaternary structures.

Fig 1. Illustrations of a/an (a) α -helix. (b) Anti-parallel β -sheets. (c) Parallel β -sheets. Figure (a) was generated by author using PyMOL.¹ Figures (b) and (c) were obtained from the Internet from Wikimedia Common Repository.^{2,3}

Central to proteins is the structure and function paradigm, in which it is proposed that the structures adopted by any protein is responsible for its biological function. Recently however, intrinsically disordered proteins (IDP), which are defined as proteins without well-defined structures and energy minima, have also received considerable interest due to its abundance in the genomic database.^{15–17} For example, Wright and Dyson have recently published a review on the roles of IDP cellular signaling and regulation.¹⁸ A protein can consist of up to thousands of amino acids, giving rise to an enormous amount of possible conformations to be adopted. This is classically known as the Levinthal's paradox, when Cyrus Levinthal first noted this problem in 1968.¹⁹ In Levinthal's paradox, every possible conformation of a sequence of amino acid is assumed to have identical weight and probability. Therefore, their relative free energies are identical. As such, as the number of amino acid increases, the conformational space grows extremely fast and Cyrus Levinthal showed that it takes an enormous amount of time to sample the conformational space in order to arrive at the native conformation.

Further discussion on the topic was able to reduce the complexity of the problem through statistical energy landscape theory.^{20–22} It was proposed that the free energy

landscapes of proteins are rugged and funnel-shaped. Due to this funnel shape, there is a force bias towards the conformations with low free energy, ultimately allowing the protein to attain the native state at the global minimum point on the free energy landscape. Moreover, we can see that it is possible for proteins to go through different pathways in order to reach the same native state depending on both its initial state and environmental factors such as temperature. The rugged-ness of free energy landscape also suggests that there are various local minima along the pathways on the free energy landscape of which the proteins may be trapped temporarily before proceeding to other conformations with probabilities depending on the free energy difference. In some cases, the free energy barrier in certain local minimum may be too high for the proteins to overcome, causing them to be “stuck” with the particular conformation belonging to the particular local minimum. The situation when proteins fail to arrive at its native functional structure is described as protein misfolding. The phenomenon has been linked to proteins diseases such as Alzheimer’s Disease, Parkinson’s Disease, mad cow disease etc.^{23,24}

1.2 Proteins' Aggregation and Gelation

Protein-protein interactions is one of the most important topics among the wide range of topics studied in biology for many reasons. Among the various protein-protein interactions, aggregation of proteins has received a lot of amount of interest due to its pathological significance. One prominent example can be found in the study of the protein amyloid beta ($A\beta$) protein, which is a peptide linked to Alzheimer's Disease.^{25,26} As mentioned in section 1.1, Alzheimer's disease is a type of protein misfolding disease. In Alzheimer's disease, deposits of $A\beta$ called the amyloid plaque are found in patients' brain. Amyloid plaque was thought to play a role in causing neurodegeneration.²⁷ These plaques have been found to contain large amount of fibril-like structures. Experimentally, it was discovered that the main secondary structure of $A\beta$ fibrils are composed of strands of parallel β -sheets arranged in a highly ordered fashion. As an example, the structure of a protofibril (short and partially formed fibril) is illustrated in Figure 2.⁴

Fig 2. Figure shows a protofibril of amyloid beta ($A\beta$) solved experimentally. Generated using PyMOL using structure 2BEG from Protein Data Bank (PDB).^{1,4}

Another example of a proteins aggregation disease is type 2 diabetes mellitus. Type 2 diabetes mellitus is caused by aggregation of misfolded amylin followed by

its deposition in pancreas, similar to that of Alzheimer's disease with A β .^{28,29} Note that amylin is also commonly known as Islet Amyloid Polypeptide (IAPP) and these nomenclatures will be used interchangeably in this thesis. Misfolded amylin aggregates into insoluble amyloid fibrils that are toxic and affects pancreatic β cells responsible for producing insulin.^{29,30} One of the challenges posed by the study of amyloid fibrils is that the nature of amyloid fibrils is insoluble and non-crystalline, making it extremely difficult to obtain detailed 3D structures of a matured fibril via conventional experimental technique such as X-ray crystallography and 1H -NMR.³¹ As a result, we shall see in Chapter 2 that it is often useful to complement experimental study with structural insights obtained using computational simulations.

In addition to its pathological implications, proteins aggregation can also result in the formation of peptides gel, which is a versatile biomaterial. Peptide hydrogel has received considerable interest in recent years due to its potential applications in various fields such as tissue engineering and drug delivery. Hydrogels are made up of polymeric network that is able to hold high water content within the network structure. The network can be chemically or physically crosslinked.³² The high water content and porous morphology are critical in allowing nutrients, drugs or waste to diffuse easily. Self-assembling peptides such as diphenylalanine (FF), silk fibroin and peptide amphiphiles have been used to create strong nanomaterials and nanodevices.^{33–35} These peptides have received considerable interest due to the current availability of structural knowledge of folding and aggregation kinetics. For example, both β -sheets and α -helices have been found to stabilize nanofibrous hydrogels.³⁶

1.3 Protein-protein Recognition

Enzymes are catalysts in living organisms used to speed up biochemical reactions by altering the free energy barrier of chemical process.³⁷ Biological enzymes are highly efficient and are vital for sustaining the majority of biological processes necessary for survival. For example, the enzyme hexokinase is responsible for phosphorylation of glucose into glucose-6-phosphate, while many types of isomerases are used by living organisms to convert chemical compounds to their isomers.^{38–40}

Besides facilitating chemical reaction involving conversion of chemical compounds, many proteins have been found to be involved in interact with other protein species to serve diverse biological functions. This will be the second type of protein-protein interactions we will examine in this thesis. There are many types of protein-protein interactions, for example homodimeric proteins, heterodimeric proteins, enzyme-antibody etc.^{41,42} Some of the most important protein-protein interactions include regulation of actin polymerization by profilin proteins, signal transduction pathway by SH2/SH3 proteins, formation of hemoglobin heterotetramer for oxygen transport.^{43–45} Here, structures of SH3 complexes and hemoglobin heterotetramer are shown as examples in Figure 3.^{5,6} The understanding of the underlying mechanisms of which proteins may assemble into complexes is extremely important for many reasons. For example, the knowledge of chemical and physical characteristics of targeted binding sites allows one to design drug with high specificity.^{46–48} Chapter 5 of this thesis will explore the interaction between profilin and formin, both of which were involved in controlling atin polymerization.

Fig 3. (a) Structure of SH3 with a proline-rich peptide (PDB code:1PRL).⁵ (b) Structure of a hemoglobin heterotetramer (PDB code:1C7C).⁶ Figures were generated by the author using PyMOL.¹

2 Methodology and Literatures Review

2.1 Molecular Dynamics

In the past decade, predicting the structures of proteins using computers has been one of the main breakthroughs in life science. This is due to the advancement in technology, unlocking extremely fast and scalable calculations previously unimagined. Using sophisticated computational algorithms, scientists have tried to predict protein structures from its primary sequence through sophisticated methods such as sampling conformations with molecular dynamics, Monte Carlo algorithm, and using statistical method such as machine learning and graph theory.^{49–52} In this thesis, the tool of our choice is molecular dynamics.

The first molecular dynamics simulation of proteins was carried out in 1977 by Karplus et al.⁵³ Since then, it has been undergoing continued refinements to become a critical tool in probing proteins at atomistic level often inaccessible by even the most sophisticated experimental instruments. One of the very early molecular dynamics simulations of bovine pancreatic trypsin inhibitor was carried out by Michael Levitt, who was awarded the Nobel Prize in Chemistry in 2013 for his contribution to the field.^{49,54} It is interesting to note that in the 1983 paper written by Michael Levitt, structure of bovine pancreatic trypsin inhibitor was simulated for just 60 ps, a timescale that would only require minutes to complete with current technology.

Besides looking at proteins at the highest resolution, the nature of molecular dynamics, where atoms undergo time evolution through the use of force field models, enable the study of proteins folding pathways, metastable states for drugs applications, self-assembly mechanism of peptides and many more.^{55–61} This is in contrast to other methods such as Monte Carlo simulation where the time evolution of dynamics is not available. Traditionally, Monte Carlo has enjoyed advantages in overcoming free energy barriers due to its use of random sampling.^{62,63} However, developments

of sophisticated enhanced sampling algorithms in the recent years such as replica exchange molecular dynamics, hamiltonian exchange molecular dynamics and metadynamics have increased the efficiency of conformational sampling using molecular dynamics.^{64–67}

Molecular dynamics are inherently computationally expensive due to the amount of interactions necessary to be computed as the number of atoms increase. Interactions between pair of atoms scale with the complexity of $O(n^2)$ where n is the number of atoms. Most of the time, molecular dynamics simulations are carried out in nanoseconds scale. Coarse-grained simulation with implicit solvent allows exploration of significantly longer timescale at the expense of accuracy. However, in 2008, Shaw et al. employed a computational cluster designed for molecular dynamics, Anton which is capable of atomistic simulations of proteins in explicit solvents in microseconds scale.⁶⁸ While impressive at its time, technological and algorithmic improvements have allowed many smaller computational clusters to perform molecular dynamics simulation at microsecond timescale.

However, specialized computational cluster such as Anton is still an order of magnitude faster than most computational clusters we can access to. For example, using Anton, Shaw et. al was able to perform simulation in the milliseconds scale on BPTI (58 amino acids long), ubiquitin (76 amino acids long) and villin headpiece (35 amino acids long).^{69–72} Ubiquitin for example, has been known to be a relatively slow-folding protein. Shaw et. al’s simulation was able to simulate the spontaneous and reversible folding of ubiquitin, otherwise unobtainable through conventional nanoseconds simulations. Besides, milliseconds MD enables dynamical study of multiple rare events in the folding process, which typically happen only in the microseconds scale. With enough statistical significance of the events happening in a folding trajectory, accurate thermodynamic quantities such as heat capacities can be calculated and free energy landscapes may be constructed.

Other than Shaw’s group, Stanley et. al recently simulated the kinase-inducible

domain (KID) for 1.7 ms in explicit solvent by using distributed GPU-computing grid.^{73,74} They were able to determine a metastable state present only in the phosphorylated KID with slow conformational kinetics. Recently, Anton 2 and MDGRAPE-4 are under development to allow even faster molecular simulations in the near future.^{75,76} To date however, atomistic simulations in milliseconds as discussed require special-purposed computational clusters such as Anton or through the use of distributed computing, both of which we are unable to access in this thesis. As such, we chose to accelerate our molecular simulations through enhanced-sampling methods such as metadynamics and replica-exchange molecular dynamics (REMD).

Armed with highly parallel computing clusters available to us through High Performance Computing Center (HPCC) in Nanyang Technological University and more recently National Supercomputing Centre of Singapore (NSCC), this thesis aims to exploit these computational resources to study various protein-protein interactions on the *microseconds* timescale. Simulations at *microseconds* timescale require computational resources running for months and hence are rare. For example, in Chapter 3, we used 224 CPUs running continuously for two months and in Chapter 4, 1536 CPUs were used for two months in the metadynamics simulations. Specifically, our computational resources allowed us to employ molecular dynamics and multiple sophisticated enhanced sampling methods to study the atomistic interactions in three topics: Cross-species protein-protein aggregation between amylin and prion protein, binding mechanism between profilin and polyproline and lastly mechanism of peptide gelation.

The fundamental working principle of molecular dynamics lies in the use of an empirical force field to calculate forces acting on atoms in any simulated system. The forces are calculated according to Newton's equation of motion:

$$\frac{d^2 \mathbf{r}_i}{dt^2} = \frac{\mathbf{F}_i}{m_i} \quad (2.1)$$

where $\mathbf{r}_i$ is the position of i^{th} atom, $\mathbf{F}_i$ is the force acting on i^{th} atom and m_i is the mass of i^{th} atom. The empirical force field is essentially the potential of the system defined as equation 2.2

$$V(\mathbf{r}) = V_{bonded} + V_{nonbonded} \quad (2.2)$$

$$\mathbf{F}_i = -\frac{\partial V}{\partial \mathbf{r}_i} \quad (2.3)$$

where $V(\mathbf{r})$ describes the potential of the simulated system as a function of the atomic coordinates $\mathbf{r}$ where $\mathbf{r}$ is the set of coordinates for N atoms $\mathbf{r}_i$ to $\mathbf{r}_N$, V_{bonded} is the bonded potentials including covalent bond-stretching, angle-bending, improper dihedrals and proper dihedrals. $V_{nonbonded}$ is the non-covalent interactions including Van der Waal's interactions (Lennard-Jones potentials) and electrostatic interactions (Coulombic potentials).⁷⁷ A typical system of equations for molecular dynamics are as follow:

$$V = \sum_{bonds} \frac{1}{2} k_b (r - r_0)^2 + \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2 + \sum_{torsions} \frac{V_n}{2} [1 + \cos(n\phi - \delta)] \quad (2.4)$$

$$+ \sum_{improper} V_{imp} + \sum_{LJ} 4\epsilon_{ij} \left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^6}{r_{ij}^6} \right) + \sum_{elec} \frac{q_i q_j}{r_{ij}}$$

where the first term describes bonds between atoms using a harmonic potential, second to fourth term describes various angles of the systems such as angle stretching and angle bending, the fifth term describes Van der Waal's forces using Lennard-Jones potential and finally the last term describes the coulombic interactions among the atoms.⁷⁸ A system of atoms is first initiated with random velocities for all atoms conforming to the Boltzmann distribution at a desired temperature. Then, forces are

calculated according to equation 2.3. Once the forces are calculated, the velocities of atoms can be evaluated by using numerical integration scheme. One of the most commonly used scheme is the Verlet integration scheme:⁷⁹

$$\begin{aligned}\mathbf{r}(t + \Delta t) &= \mathbf{r}(t) + \Delta t \mathbf{v} + \frac{\Delta t^2}{2m} \mathbf{F}(t) \\ \mathbf{v}(t + \Delta t) &= \mathbf{v}(t) + \frac{\Delta t}{2m} [\mathbf{F}(t) + \mathbf{F}(t + \Delta t)]\end{aligned}\tag{2.5}$$

where $\mathbf{r}$ is the coordinates of atoms, $\mathbf{v}$ is the velocities of atoms, m is the mass of atoms and F is the force acting on atoms.

In the following sections we will discuss the use of molecular dynamics in studying proteins aggregation and protein-protein recognition. Thorough review of both experimental and computational works of the topics covered in this thesis is not possible, but the following sections will place their focus on providing an overview on computational studies of proteins relevant to this thesis whenever possible. In the first part while reviewing literature for proteins aggregation, we concentrate on studies on amylin and prion proteins. Then, reviews on peptides gel formed by amphiphiles and silk fibroin proteins are provided. In this thesis, the peptides used in Chapter 4 are highly similar to silk fibroin proteins. Finally, we reviewed literature on polyprolines recognition by SH3 and Ena-VASP proteins which would provide a background to our study on polyprolines recognition by profilin. However, note that comparison against experimental studies will be discussed consistently during the discussion of the individual projects in this thesis.

2.2 Applications of Molecular Dynamics on Protein-protein Interactions

2.2.1 Amylin (IAPP) Aggregation

Molecular dynamics have been used to complement experimental studies on proteins aggregation for a long time. As mentioned previously, $A\beta$ is one of the most well-studied proteins because of its pathological significance. In fact, amyloid fibrils may refer to fibrillar structures formed by other protein species similar to those of $A\beta$. Naturally, the prevalence of amyloid fibrils extends into computational studies using molecular dynamics. For example, molecular dynamics have provided extensive insights into formation of amyloid oligomer and fibrils. Generally, ab-initio folding simulations of peptides are used to sample possible stable and metastable conformations adopted by oligomers while simulations of amyloid protofibrils or proposed model are used to investigate the energetics of tertiary or quarternary aggregated proteins complexes. The study of amyloid fibrils is central to the Chapter 3 and 4 of this thesis. As mentioned before, in Chapter 3 we will investigate the dimerization process between amylin (IAPP) and a fragment of prion protein from residue 106-126 (PrP106-126). Incidentally, both amylin and PrP106-126 have been found to form amyloid fibrils in experiment similar to $A\beta$.⁸⁰⁻⁸²

Around 110 years ago, hyaline lesion was discovered in islets of Langerhan by Opie.⁸³ In the beginning, this lesion was thought to be composed of insulin deposits. However, in 1987, Westermarck et al. and Cooper et al. characterized the amylin fibril contained in these deposits in the hyaline lesion.^{84,85} Amylin consists of 37 residues. It is found in the vicinity of insulin in β -cell secretory granules. The crystal structure of amylin has been solved by Nanga et al. and is illustrated in Figure 4.⁷ The amino acids region from residue 20 to residue 29 on amylin is known to be highly amyloidogenic.^{86,87} Figure 4 shows an experimental structure of amylin fibril formed by hIAPP20-29 (SNNFGAILS) solved by Nielsen et al. using solid state NMR (nuclear

magnetic resonance spectroscopy).⁸ In a study by Madine et al., they proposed both parallel and anti-parallel β -strands structures using data from both solid-state NMR and X-ray diffraction.⁸⁸

Fig 4. (a) Crystal structure of full length human amylin (hIAPP). (b) NMR structure of anti-parallel β -sheets structures of hIAPP20-29 (SNNFGAILS). Figures are generated by author using PyMOL.^{1,7,8}

There have been numerous computational studies attempting to provide valuable insights into structural characteristics of amylin fibril. Cecchini et al. used both implicit and explicit solvent molecular dynamics to look at the structural determinants of forming amyloid fibrils.⁸⁹ Both A β and amylin were studied in the paper. The author employed a novel methodology by simulating the interested proteins in small overlapping fragments to obtain good conformational sampling from each region. Then, the results from all segments were combined and extrapolated into the amyloidogenic and secondary structure profiles of the full peptides. In that paper, the author was able to determine two hotspots favorable for β -sheets formation: residue 10-22 and residue 28-30. This was consistent with experimental study by Gazit et al. using in vitro immunoblotting.⁹⁰ Besides, Cecchini et al.'s simulation suggested that there were two short segments that adopt turn-like structure: GLY24 and ALA25, GLY33 and SER34, both of which were able to regulate the β -sheet profiles of amylin. This regulation of

β -sheets profile may have an effect on the aggregation propensity of amylin.

In addition to that, Laghaei et al. used hamiltonian replica exchange to study the dimerization of human and rat amylin.⁹¹ Hamiltonian replica exchange is a variant of molecular dynamics that has been used to accelerate conformational sampling in protein folding problem.^{65,66} Other than studying dimerization, The role of disulfide bridge formed between residues CYS2 and CYS7 was also investigated. This disulfide bridge has been shown to stabilize α -helical conformations at the N-terminal region of amylin.⁹² Furthermore, their secondary structure analysis using their simulations results were consistent with experiment in that region 20-29 was found to adopt β -sheets. In fact, it was found that the amyloidogenic region may extend down to residue VAL17. Secondly, the N-terminal region was able to fold into α -helix. N-terminal α -helix was found to play a vital role in anchoring amylin to lipid membranes.⁹³

Other than that, a recent study by Liang et al. in 2013 used explicit solvent molecular dynamics to understand the difference between amylin present in humans and in rat.⁹⁴ Rat amylin, when overexpressed, do not cause rats to develop diabetes-like symptoms. Interestingly, there are only differences in six amino acids when comparing between rat amylin and human amylin. In Liang et al.'s paper, besides simulating the monomer, parallel β -sheets protofibrillar models were built and simulated to test the stability of the models. In simulations of oligomers, it was found that human amylin was able to maintain higher stability with increasing number of peptides, while rat amylin were comparatively unstable in the corresponding oligomeric configurations. On the other hand, simulation of monomers showed that both human and rat IAPP were relatively unstable and had high structural polymorphism.

2.2.2 Prion Fragment: PrP106-126 Aggregation

The major prion protein (PrP) is a protein found mostly in the nervous system of humans. The full length PrP is 253 residues long. While its physiological functions remain unclear, PrP has been linked to various neurodegenerative diseases such as

mad cow disease, Creutzfeldt-Jakob disease and Gerstmann-Sträussler syndrome.^{95,96} Generally, human PrP can be characterized by two segments; Residue 23-125 at the N-termini, which is intrinsically disordered and thus lacks well-defined secondary structure. On the other hand, the C-terminal region has well-defined structure as shown in Figure 5(a).⁹

Stanley Prusiner was awarded the Nobel Prize in Physiology or Medicine in 1997 for his research into PrP. It was found that the main pathological pathway of PrP is through the conversion of the α -helix rich PrP^C into the β -sheet rich PrP^{SC}.⁹⁷⁻¹⁰¹ PrP^{SC} forms amyloid fibril-like structures similar to A β and amylin. Knaus et al. solved the crystal structure of dimeric PrP which was thought to be an important factor in the conversion of PrP^C into PrP^{SC} using X-ray diffraction as shown in Figure 5(b).¹⁰ Sekijima et al. simulated both the monomeric and dimeric form of PrP using explicit solvent molecular dynamics.¹⁰² It was found that while both monomeric and dimeric PrP had similar dynamics, the dimeric form was able to maintain its tertiary structure better than that of the monomeric PrP. DeMarco and Daggett used molecular dynamics to create a profibril model containing β -sheet rich structures consistent with experiment.¹⁰³ In their paper, Syrian hamster PrP was simulated with a mutation known to cause conversion into PrP^{SC} at low pH. Thereafter, a prion aggregate model consistent with electron microscopy (EM) data from Speare et al.¹⁰⁴ was built using structures obtained from the simulation.

Fig 5. (a) Solution NMR structure of human prion protein (PDB Code:1QLX).⁹ Structures shown are for residue 125 to residue 228. (b) X-ray crystal structure of dimeric prion.¹⁰ Figure generated by author using PyMOL.¹

Due to its long amino acid sequences, ab-initio folding simulation of complete PrP is extremely difficult as the conformational space is enormous. A common approach to studying large proteins such as PrP is to look for short segments extracted from the protein itself. The short segments however, should exhibit similar characteristics to that of the complete protein. In this context, a small fragment from residue 106 to residue 126 of PrP, or PrP106-126 in short was discovered to have similar neurotoxicity to that of PrP^{SC}.^{105–107} Levy et al. carried out one of the first simulations on PrP106-126 using both implicit and short explicit solvent to investigate the structural stability of α -helix and β -sheets.¹⁰⁸ It was found that HIS111 and VAL122 played important roles in helix-coil transition of PrP106-126. Other than that, α -helix was found to be unstable in neutral condition.

Villa et al. simulated PrP106-126 under four different solvents: water, dimethyl sulfoxide (DMSO), trifluoroethanol (TFE) and hexane.¹⁰⁹ Simulations were carried out from 10 ns to 60 ns. In agreement with experiments, it was found that the conformations of PrP106-126 depends greatly on the solvents. Under physiological pH in water, PrP106-126 was able to form significantly more β -sheets than those other solvents. Furthermore, α - β transition was observed under the same condition.

Apart from that, Grabenauer et al. ran replica-exchange molecular dynamics (REMD) simulation to complement their ion mobility spectrometry-mass spectrometry (IMS-MS) experiments on PrP106-126.¹¹⁰ In that paper, monomeric PrP106-126 was found to form β -hairpin structures in implicit solvent REMD simulation in agreement with the spectrometry results. It was also suggested that the oligomerization of PrP106-126 may result from the assembly of the monomeric β -hairpin structures.

A more recent molecular dynamics study of PrP106-126 was carried out by Ning et al. in 2014.¹¹¹ Extensive implicit replica exchange molecular dynamics (REMD) was employed in their study. Similar to hamiltonian replica exchange, REMD is used to accelerate conformational sampling in molecular dynamics.⁶⁴ Simulations was carried out on both monomeric and dimeric PrP106-126. Moreover, long conventional molecular dynamics simulation using implicit solvent was carried out on PrP106-126 trimer and tetramer formation. Both monomeric and dimeric PrP106-126 exhibited high structural polymorphism, in agreement to the earlier study by Villa et al.¹⁰⁹ MD simulation of trimer and tetramer found stable conformations with high β -sheet content.

2.2.3 Peptides Gelation

The association of peptides into hydrogel is also an aggregation phenomenon. Hydrogels are water-swollen polymeric material and they are typically characterised by well-defined 3D network of polymers.¹¹² In the case of peptide hydrogels, the polymeric network is made up of the supramolecular assembly of peptides. Gazit's group has carried out extensive characterization of peptides hydrogel formed by simple diphenylalanine analogs.^{113–115} Remarkably, in one of their studies, a hydrogel formed with just 1 % of peptide materials was able to maintain its 3D volume well.¹¹³ Microscopic images showed that the gel networks were formed by amyloid-like fibrils and Fourier transform infrared (FTIR) data showed characteristic spectral peaks of β -sheets. Another one such commonly found peptide hydrogel is formed by amphiphilic peptide

(PA), which is a type of molecule designed to self-assemble into nanofibers.¹¹⁶ Peptide amphiphiles (PA) consist of lipid molecules attached to a sequence of hydrophilic peptides, allowing it to arrange into highly ordered networks of fibrils, nanotubes or micelles. Interestingly, β -sheet has been suggested as a crucial factor in the self-assembly of PA.^{117–119}

The mechanism in which peptide amphiphile arranges into ordered networks starts with the aligning of the hydrophilic (can form hydrogen bonds) amino acids towards the surrounding water. This was followed by burying the hydrophobic lipid molecules inside. This type of peptide nanofibers has been simulated using molecular dynamics. In Lee et al.'s simulation, self assembly of 144 PAs was simulated using explicit solvent molecular dynamics.¹²⁰ The PAs consist of an alkyl chain attached to peptide sequence **SLSLAAAEIKVAV**. PAs were initiated in a nanofiber-like arrangement and it was found to be stable after 40 ns of simulation. Notably, the β -sheet population found in the simulation is consistent with experimental results using circular dichroism study. A recent paper by Tekin studies self assembly of another type of PA molecule consisting of an alkyl chain attached to a peptide sequence **VVAGERGD**.¹²¹ Similar findings to Lee et al. were made in that the constructed nanofibers were stable and β -sheets were found to play important roles in stabilizing the nanostructures. Velichko et al. found the interplay between hydrophobic interactions of alkyl chain and hydrogen bonds between peptides to be important in regulating the morphology of peptide assemblies.¹²²

Besides peptide amphiphiles, some hydrogels can be formed by the self-assembly of amyloid fibril such as those formed by α -synuclein proteins.^{123,124} Peptide gels using naturally occurring peptides provide an inherent advantage in that they do not require the synthesis of artificial chemical compound unlike peptide amphiphiles. A well known protein that forms hydrogel is the silk fibroin protein. Silk is a type of polymeric protein produced by silkworms and spiders in fiber form.¹²⁵ The most well-characterized protein from silk protein is the silk fibroin protein from *Bombyx*

mori. Silk fibroin protein is known to be a major component that contributes to the formation of β -strands in silk fibers.^{33,126} The crystal structure of N-terminal domain of fibroin was solved by He et al. recently in 2012 and its β -sheet rich structure is shown in figure 6.¹¹

Molecular dynamics study was carried out by Yamane et al. to study silk fibroin proteins.¹²⁷ In this paper, model dipeptide Ac-XXX-NHMe where XXX is GLY, ALA or SER was simulated in addition to a longer Ac-(ALA-GLY)₈-NHMe Alanine-Glycine repeat. While the simulations were short (likely due to limited computational power in year 2002), the study found a high possibility of formation of β -turn in the longer Alanine-Glycine repeat. A more recent study by Dubey et al. looked into the effect of calcium ion on promoting gelation of fibroin protein.¹²⁸ Interestingly, a predominantly α -helical intermediate conformation was found to persist in their experiment and molecular dynamics was used to pinpoint the locations in which interactions between calcium ions and amino acids on fibroin proteins take place.

Fig 6. Crystal structure of the N-terminal domain of silk fibroin from *Bombyx mori*. Figure generated by author using PyMOL.^{1,11}

Recently, a novel methodology in computational simulation of peptide gel was

proposed by Frederix et al.^{129,130} They employed Martini model, a coarse-grained force field used to reduce a complex molecular systems into simple beads in order to accelerate sampling of conformations.^{131,132} In Frederix et al.'s study, all possible combinations (8000) of tripeptide sequences using all amino acids were simulated using Martini model for 50 ns. Then, the peptides that showed high aggregation propensity were simulated for a longer time to observe transition into supramolecular structures. Next, experiments were carried out on a few chosen peptides that showed promising structural features. The predictive power of Martini model was supported by the discovery of four tripeptides (KYF, KYY, KFF and KYW) that were able to form gel. In Chapter 4, we will follow similar approach in studying gelation of a short peptide sucker ring teeth (SRT) proteins. Detailed explanations of the Martini model will be provided in the later Chapter.

2.2.4 Protein-protein Complexes: Recognition of Polyprolines

In this thesis, protein-protein recognition refers to the interactions between proteins that result in formation of biologically functional quaternary complexes. In Chapter 5, we will simulate interactions between profilin and a short polyprolines sequence from FH1 domain of formin proteins. As literature on the computational study of profilin is lacking thus far, our study will fill in an important gap on research in this area. Fortunately, literature exist on molecular dynamics simulations of other protein-polyproline complex, which will shed lights into the mechanism of polyprolines recognition by proteins receptors. An example of SH3 in complex with a polyprolines ligand was shown in Figure 3(a) in Chapter 1.3.

Besides SH3, Ena-VASP is a cytoskeletal protein and it is one of the main components at the signal transduction pathway of human platelets.¹³³ The EVH1 domain was found to recognize polyprolines with the sequence motifs of **"FPPPP"**.¹³⁴ Ball et al. carried out NMR spectroscopy to characterize the recognition of polyprolines with **"FPPPP"** motifs by EVH1 domain of the VASP proteins. Molecular dynamics was

used to simulate a model of EVH1 in complex with the polyprolines ligand.¹³⁵ The model was built based on the chemical shift results obtained from NMR and it was found that the polyprolines recognition was mainly driven by the hydrophobic residues present on both receptors and peptide. The atomistic details allowed the author to explain various chemical shifts present in the NMR data.

In a recent study by Hou et al., molecular dynamics simulation was carried on Abl SH3 domain in complex with 35 ligands including both binders and non-binders.¹³⁶ The study aimed to elucidate the structural determinants of sequence motifs recognized by the SH3 domain. MM/PBSA protocol was used to estimate the binding free energies of each of the complexes simulated in the paper. The relative ranking of binding energies correlated well with experiments and was able to distinguish between the binders and the non-binders. Both Van der Waal's and electrostatic interactions were found to be crucial in the peptide recognition mechanism. Computational mutagenesis was carried out and by using a novel approach where the author was able to pick 10 out of 13 binding partners of Abl SH3 domain directly from the human genome. One of the shortcomings of MM/PBSA analysis is its poor accuracy in predicting the absolute binding free energy.¹³⁷ Besides that, another interesting study was carried out by Gumbart et al. using sophisticated method involving alchemical transformation and potential mean force to calculate the absolute binding affinity of Abl SH3 in complex with a polyprolines ligand.¹³⁸ The results were found to be in excellent agreement with experimental values.

2.3 Aims and Objectives

This thesis consists of three major works carried out in exploring protein-protein interactions covering the topics introduced in Chapter 1. In Chapter 3, we study the dimerization of amylin and PrP106-126. This was motivated by the Gal et al.'s experiment in which the two proteins were found to cross-fibrillate.¹³⁹ Typically, proteins aggregation happens between those of the same proteins species, as can be seen from both the examples of A β and IAPP mentioned in Chapter 1.2. This is mainly due to the fact that proteins' amyloidogenicity are highly dependent on amino acid sequences. Hence, when amyloidogenic proteins such as A β and IAPP come together, it is highly feasible for amyloid fibril formation.

In the past decade, the phenomenon of aggregation between proteins of different species has generated good research interest as well. It was discovered that patients with diabetes mellitus has increased risk factor of dementia in one study by Ott et al.¹⁴⁰ This implies that there can be pathological significance for cross-species proteins aggregation. This was supported by a paper in which it was found that IAPP and A β may cross seed amyloid fibrils formation.¹⁴¹ In Chapter 3, we also explored the aggregation mechanism between IAPP and PrP106-126, two sequentially and structurally different peptides. Cross fibrillation phenomenon is a relatively new field compared to the study of amyloid fibrils. Thus, our work provides insights into the early stage of aggregation between two highly different peptides. This work was carried out on the Institute of High Performance Computing Center in Nanyang Technological University. This work has been published in 2016.¹⁴²

Next, in Chapter 4, we employ a novel approach in which both coarse-grained and atomistic simulations of a short peptide sequence extracted from the sucker ring teeth (SRT) protein from squid. This work in Chapter 4 was carried out in collaboration with Assoc. Prof. Ali Miserez. Out of two highly similar peptides, only one was able to form peptide hydrogel. We performed coarse-grained modelling using Martini force

field to investigate the supramolecular assembly of the two peptides. Then, extensive molecular dynamics simulation using enhanced sampling methods was carried out to fold the tetramers of the two peptides. This allowed us to examine the self-assembly process on two different scales. The recently established National Supercomputing Centre of Singapore allowed us to carry out large scale simulations in parallel for very long timescale ($\approx 20\mu s$) even with atomistic modelling, providing a valuable study into the self-assembly mechanism of peptide gel. This is still an ongoing project with more experimental data expected in the near future.

The last work in this thesis was carried out in collaboration with Asst. Prof. Miao Yansong from the School of Biological Sciences. In Chapter 5, we study polyproline recognition mechanism by profilin from plant species *Arabidopsis thaliana* (AtPRF). Using conventional molecular dynamics and MM/PBSA analysis, we constructed a model of AtPRF in complex with polyprolines. Note that such a structure could not be obtained from crystallography. The unique findings of this work include the effect of electrostatic interactions and CH- π interactions in the recognition mechanism. This work has been included in a manuscript in preparation for submission. In addition to that, we had also studied the interactions between a profilin species containing extended N-termini with numerous positively-charged amino acids and polyprolines. The computational data was found to be in good agreement to experimental data from Asst. Prof. Miao Yansong (Manuscript in preparation¹⁴³).

3 Replica Exchange Molecular Dynamics Simulation of Cross-Fibrillation of IAPP and PrP106-126

3.1 Introduction

Cross-fibrillation phenomenon of proteins has been an area of research gaining increasing interests, mainly because of the observations that patients afflicted with certain amyloid disease can acquire unrelated protein misfolding disease. For example, patients with Type II diabetes are associated with elevated risk of dementia.¹⁴⁰ Indeed, a recent study found that IAPP and amyloid- β ($A\beta$) can cross-fibrillate.¹⁴¹ The fibrillation however, may be due to sequence similarity (37% identity evaluated using BLAST¹⁴⁴) between them. In another experiment, co-localization of prion protein and $A\beta$ was observed in the form of amyloid plaque containing fibril structures.¹⁴⁵ Both studies imply that cross-fibrillation between IAPP and prion protein could be a favorable process.

Recently, N. Gal et al. studied cross-fibrillation between two distinctly different amyloidogenic protein species that are the human amylin (Islet Amyloid Polypeptide, IAPP) and the PrP106-126 fragment.¹³⁹ In the experiment, ThT fluorescence implied β -sheet secondary structure formation between the two species. Other than that, it was found that the PC/PG lipid bilayer used in the experiments significantly enhanced the rate of aggregation between the two proteins species. It was suggested in the study that the lipid bilayer may increase the local concentration of the proteins, thus reducing the degree of freedom for the interaction between the two proteins, subsequently resulting in the increased rate of aggregation. The molecular level mechanism of this cross-fibrillation phenomenon was however, not well understood. It is the aim of the current study to explore the cross-fibrillation mechanism by examining a simple model of

dimerization of the two different peptides, IAPP and PrP106-126. In our study, we mimic the high local concentration of proteins in the experiment by choosing the simulation box size such that the concentration of the peptides is roughly 3.7 mM.

In this chapter, we have investigated into the aggregation process of a prion fragment, PrP106-126 and IAPP using extensive replica exchange molecular dynamics (REMD) simulations. REMD simulation is an enhanced sampling variant of molecular dynamics simulations.⁶⁴ IAPP and PrP106-126 do not share (again verified by BLAST) sequences (refer to Table 2) or structural similarity, hence it would be interesting to study how these two unrelated proteins may aggregate. Through REMD simulations, we hope to elucidate at the atomic level the dimerization mechanism of PrP106-126 fragment and IAPP. Dimerization process has been shown to be important in the early stages of proteins fibrillation process.^{146–148}

K	T	N	M	K	H	M	A	G	A	A	A	A	G	A	V	V	G	G	L	G
106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126
K	C	N	T	A	T	C	A	T	Q	R	L	A	N	F	L	V	H	S	S	N
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
N	F	G	A	I	L	S	S	T	N	V	G	S	N	T	Y					
22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37					

Table 2. Assignment of index to residues. First row is the indexes for PrP106-126 while the second and third rows contain indexes for IAPP. Note that index 22 is reserved as a chain separator to ease visualization.

3.2 Simulation Details and Analysis Methods

Natively amidated IAPP structure was obtained from RCSB Protein Data Bank, code 2L86, which is the NMR structure of IAPP resolved in a membrane environment.⁷ PrP106-126 fragment, an intrinsically disordered sequence, was constructed using SwissPDB Viewer into an uncapped linear strand.¹⁴⁹ To facilitate the analysis, we labelled the residues by assigning indexes as shown in table 2. The OPLS-AA force field was used.⁶³ The peptides were put into a dodecahedron water box containing 7025 SPC/E water ($\approx 1000 \text{ kg m}^{-3}$) molecules. Periodic boundary conditions were applied on the three dimensions of x, y and z. The initial box volume was $\approx 222 \text{ nm}^3$. The initial inter-chain distance was set at $\sim 7 \text{ \AA}$. Charges were neutralized by using 20 Na^+ and 25 Cl^- ions which also maintain an ionic concentration of 0.15 M. GROMACS 4.6.3 was used to perform the simulation.¹⁵⁰ LINCS protocol was used to constrain bonds involving the hydrogen atoms to allow integration time step of 2 fs.¹⁵¹ Particle Mesh Ewald with cut-off of 0.9 nm was used.¹⁵² Pressure coupling was regulated using the V-rescale thermostat.¹⁵³ 1.2 nm cut-off was used for vdW interactions. Coordinates were saved every 1 ps.

Replica Exchange Molecular Dynamics (REMD) is an enhanced sampling method to overcome limited sampling of conventional MD. Copies of identical system are evolved concurrently at different temperatures. After every k^{th} steps, the replicas attempt to exchange temperatures between each other by using the Metropolis criterion in order to satisfy detailed balance condition.⁶⁴ The exchange probability is defined as:

$$P = \exp[(\beta_i - \beta_j)(E_i - E_j)], \quad (3.1)$$

where E_i is the potential energy of replica i at temperature T_i and $\beta_i = 1/(k_B T_i)$. In our simulation, 56 replicas of the system were created. The set of 56 temperatures from 315 K to 450 K was generated using the T-REMD web server to ensure the exchange

probabilities for all ensembles were more than 20%.¹⁵⁴ In each replica the proteins were rotated in both polar and azimuthal angles with respect to the principal axes of the individual proteins in order to speed up the convergence of the simulation. We are mainly interested in the slightly higher but still physiological relevant temperature of 315 K. Higher temperature was chosen to sample wider phase space in a shorter simulation time. Exchanges were attempted every 2 ps. The replicas were simulated in parallel for 600 ns, resulting in an effective simulation time of 33.6 μ s.

For simulation data analysis, inter-residues distance map was generated using GROMACS utility `g_mdmat`.¹⁵⁰ To carry out dihedral principal component analysis (dPCA),¹⁵⁵ we employed built-in GROMACS utilities `g_angles`, `g_covar`, `g_anaeig` and `g_sham` to perform the eigenvectors and eigenvalues determination and matrix diagonalization.¹⁵⁰ For clustering analysis we employed built-in `g_cluster` utility of GROMACS.¹⁵⁰ Estimated binding free energy was calculated using MM/PBSA method by applying the `g_mmpbsa` utility developed recently by Kumari et al.¹⁵⁶ SASA only model was used to estimate the non-polar solvation energy. Bootstrap analysis was used (2000 steps) to estimate the standard errors. Parameters used for `g_mmpbsa` are provided in the supporting information. Except for REMD diagnostic tests conducted in the following section, data analysis was performed by discarding the initial 100 ns of the ensemble to avoid potential statistical bias introduced during the initial period of equilibration. Figures of proteins were generated with PyMOL 1.7.4.¹

Note that in our MM/PBSA calculation, entropy contribution was ignored. This is because in MD there is currently no reliable approach to calculate entropy. Quasi-harmonic approach (which we will use in section 3.3.1) for example, only provides an upper bound for the entropy. Readers are encouraged to refer to the papers by Genheden et. al and Hou et. al on the effects and errors of entropy calculations.^{137, 157} In a nutshell, entropy calculations can be very important in trying to obtain the absolute binding free energy. However, it provides satisfactory accuracy in ranking relative binding affinities. It is also useful to use MM/PBSA to find out the important

contributions from different types of interactions in protein-protein binding, which is the main purpose behind using MM/PBSA in this thesis.

3.3 Results

3.3.1 REMD Diagnostic Tests

For an efficient REMD simulation, the set of temperatures chosen should guarantee sufficient overlap between the potential energy distributions for each temperature. This is to ensure effective exchange between replicas.^{158–160} We plotted the potential energy distribution for all temperatures in Figure 7(A) and found sufficient overlap between all temperatures. Furthermore, Figure 7(B) shows that random walks in the temperature space have been achieved for all replicas, hence we deduce that the REMD exchanges were effective. In addition, we cut the full trajectory into multiple time window of 50 ns and calculated the average structural propensity in each time window for coil and β -sheet. Ideally, the average propensity should fluctuate little across the different time windows for a fully converged simulation. The average propensities for coil (Definitions of coil is defined by DSSP) show little changes in the different time evolution windows as shown in Figure 7(C). Moreover, Figure 7(D) illustrates reasonable convergence for β -sheet propensities, in particular after 300 ns. In Figure 7(E), the time evolutions of various secondary structures under 315 K are shown. The secondary structures sampled show small fluctuations with respect to the means, a behavior that should be expected for good sampling in an REMD simulation. All secondary structures in Figures 7(C), (D) and (E) were calculated using DSSP algorithm.¹⁶¹ Finally, Figure 7 (F) showed that the entropy of the protein complex in bound state estimated using the quasi-harmonic approach maximizes and stays relatively constant towards the end of the simulation. The entropy was estimated using the backbone of the protein complex.

Fig 7. Test for validity of REMD simulation. (A) Sufficient overlaps are found between the potential energy distribution of all the temperatures used in the simulation. Note that each color represents the distribution from a single temperature. (B) Random walk in the temperature space for all replicas (y-axis). Colors indicate temperature ensembles, ranging from lowest temperature (blue) to highest temperature (red). (C) The average coil propensity at different temperatures are similar for ten periods during the simulation. (D) Average turn propensity which has been evaluated in a similar fashion as (C). (E) Evolution of secondary structures for the 315 K ensemble. (F) An estimate of the conformational entropy of the protein complex in bound state by the quasi-harmonic method.

3.3.2 Secondary Structures Propensity and Contact's Probability

To evaluate conformational preference of PrP-IAPP dimer, we plotted the secondary structure propensities of all the residues as shown in Figure 8. Most residues mainly adopt turn and coil. The propensity of β -sheet peaks around 0.2 for PrP106-126.

The high β -sheet propensity regions of PrP106-126 reside close to the N-terminus and C-terminus, implying a tendency to form β -hairpin; whereas for IAPP, several segments have roughly similar β -sheet propensity ~ 0.1 , with Thr30 and Asn31 peaking at ~ 0.2 . PrP106-126 has almost no tendency to form helical structures, whereas IAPP adopt α -helix close to the N-terminus at region Cys7-Ala13. Interestingly, the simulation shows that IAPP has relatively high propensity to form the more compact 3_{10} helix at the region of Thr4-Leu16. The results are consistent with the previous experimental studies, where NMR measurements show sampling of helical structure at the N-terminus of IAPP.^{7,162} 3_{10} -helix was shown to be adopted in several membrane proteins despite its lower stability compared to α -helix.¹⁶³

Fig 8. Propensity of the different secondary structures adopted by PrP106-126 and IAPP during the 100-600 ns period. The gray bar indicated by CS separates the two chains. Index 106-126 is PrP106-126 while index 1-37 is IAPP.

To illustrate the interaction pattern between the two peptides, the contacts between residues formed by heavy atoms were investigated. A contact is defined by a heavy

atom pair between the two peptides having a cut-off distance of less than 0.5 nm. The contact frequency was normalized and plotted using bar plots as shown in Figures 9(A) and (B). PrP106-126 tends to make contacts with IAPP frequently through residues Asn108 to Met112 located close to N-terminus or residues Val121, Val122, and Leu125 located close to the C-terminus. These residues may play a role in helix destabilization of IAPP preceding the transition into β -sheets. On the other hand, IAPP tends to make contacts with PrP106-126 through several segments except for the N-terminal region. Among them, residues Arg11, Phe15 and Tyr37 are found to have extensive contacts with PrP106-126.

To further understand the contact pattern of the two peptides, the inter-residues distance map was plotted. The mean smallest distance between residues was calculated using only the heavy atoms. Figure 9 shows the inter-residues distance map. As expected, the inter-residues distance map is generally consistent with the contact probability bar plots with respect to the contact hotspots. There are two hot spots on the y-axis corresponding the N- (index 109-112) and C- (index 121-125) terminal regions of PrP106-126. On the other hand, looking at the x-axis, IAPP is found to have three hot spots, each in relatively close distance with the two hot spots of PrP106-126 respectively. In particular, the closest distances can be found between 10 pairs of residues, Gly114-Leu12, Gly114-Phe15, Gly114-Ile26, Gly114-Leu27, Met112-Phe15, Met109-Asn14, Met109-Phe15, Met109-Leu16, Met109-Val17 and Met109-Leu27. Three out of the ten pairs involve the aromatic phenylalanine residues. Clearly these results show that aromatic-aliphatic interactions are very important. A small region of high contacts is observed between residues 116 and 117 which belong to the palindromic region AGAAAAGA of PrP106-126 and residues Ala25-Leu27 which belong to the important SNNFGAIL (Residue 20-27) region of IAPP. Both of these regions have been shown to play vital roles in homogeneous oligomer formation of the individual peptides.^{82, 110, 111, 164, 165}

Fig 9. Frequency of contacts formed by (A) PrP106-126 towards IAPP and (B) IAPP towards PrP106-126. (C) Mean distance between residues of IAPP and PrP106-126. All contacts and distances were calculated by including only the heavy atoms.

3.3.3 Dihedral Principal Component Analysis

In order to characterize the structural ensemble sampled during the REMD simulation, we employed the method of dihedral principal component analysis (dPCA) developed by Mu et al.¹⁵⁵ to estimate the free energy landscape of the studied system at 315 K. The dPCA projection was separated into 100 by 100 bins for the first and second principal components respectively. The frequency of projection in each bin was then

calculated. Subsequently, the relative free energy value was determined using the formula $E_i = -RT\ln(P_i/P_0)$ where R is gas constant, T is temperature (315 K). P_i is the population of each bins, while P_0 is the population of the most populated bin. The population matrix was then mapped onto a contour plot. The energy landscape for our simulation is shown in Figure 10. To find the representative structures of each local minimum, we carried out clustering at each minimum using the GROMOS algorithm with the *g_cluster* utility of GROMACS 4.6. The binding energies estimated by MMPBSA is tabulated in Table 3.

Fig 10. Contour plot of the relative free energy landscape determined from dPCA using the first and second principal components of REMD simulation at 315K. Note that the color bar represents the relative free energy (kJ mol^{-1}), which is evaluated from the population of conformations obtained by determining the frequency of the conformation within a grid of the energy landscape.

Fig 11. (A-F) Representative structures for six minima in the dPCA relative free energy landscape. Note that red color is used for IAPP, while blue color is used for PrP106-126. Yellow spheres indicate the N-termini. Figures were generated with PyMOL. (G-H) Representative structure from the 9th cluster determined from conformational clustering. (I) Secondary structures determined by DSSP for A-G. Z is the initial configuration of the system. Yellow indicates β -sheets, gold indicates β -bridge, orange indicates 3-10 helix and cyan indicates α -helix.

	Population	ΔE_{elec}	ΔE_{vdW}	ΔG_{polar}	$\Delta G_{nonpolar}$	$\Delta G_{binding}$
A	8.4%	-37.2 ± 0.4	-89.7 ± 0.2	109.7 ± 0.4	-10.8 ± 0.0	-28.0 ± 0.3
B	5.9%	14.3 ± 1.2	-79.5 ± 0.3	51.5 ± 0.5	-8.9 ± 0.0	-22.6 ± 0.3
C	1.5%	-17.3 ± 0.8	-123.2 ± 0.8	101.8 ± 0.8	-12.8 ± 0.1	-51.5 ± 0.7
D	1.3%	24.7 ± 1.0	-56.1 ± 0.7	38.7 ± 1.0	-6.0 ± 0.1	1.3 ± 0.8
E	3.5%	-18.1 ± 1.0	-117.3 ± 0.9	89.5 ± 1.0	-12.1 ± 0.1	-57.9 ± 0.9
F	4.6%	9.2 ± 0.9	-109.4 ± 0.6	72.5 ± 0.8	-11.1 ± 0.1	-38.7 ± 0.6
G	2.2%	6.3 ± 0.7	-61.0 ± 0.7	47.4 ± 0.9	-7.0 ± 0.1	-14.3 ± 0.6
H	3.6%	1.9 ± 0.8	-75.5 ± 0.9	73.4 ± 0.9	-8.6 ± 0.0	-8.8 ± 0.6

Table 3. Binding free energy estimated using MM/PBSA approach. The energies are rounded to one decimal place and shown in kilocalories per mol. The binding energy is decomposed into individual component by g_mmpbsa where ΔE_{elec} is the molecular mechanics (MM) electrostatic energy, ΔE_{vdW} is the MM van der Waals energy, ΔG_{polar} is the polar solvation energy and $\Delta G_{nonpolar}$ is the non-polar solvation energy.

It should be noted that there are two limitations found in the dPCA method in our study. Firstly, multiple structures may be projected to the same minimum. Secondly, very similar structure may be projected to separate minima, albeit they should be in close proximity to one another (Figure S1 in Supporting Information). Hence, the depth of the minima indicated by deeper color in Figure 10 should be interpreted very carefully. While attempting to identify the meaningful minima, the aforementioned issues are taken into consideration by ensuring that the representative structures populate more than 80% of the structures located in the grids corresponding to the minima. Subsequently, six well separated energy minima were identified. The representative structures of each minimum are shown in Figure 11. The percentage populations P tabulated in Table 3 were calculated by first plotting the RMSD of the full 500 ns trajectories using Structures A-F as the references. Then, the fraction of RMSD less than 0.25 nm was tabulated as percentage population P for each individual structure. We have also attempted dynamical reweighting using Multistate Bennett Acceptance Ratio (MBAR) as described by Chodera et. al and found the energy diagram (Figure S4 in supporting information) to be similar.^{166,167}

Figure 11 shows the structures from the six minima. Structure B, C, E and F show that IAPP and PrP106-126 can indeed form proteins complex with inter-chain β -sheets. Moreover, structure E contains a very high percentage of anti-parallel β -sheets.

This interesting structure hints at the possibility of further elongation into extensive fibrils. However, the global minimum, structure A, consists mostly of random coils and turns, except for a short segments of β -sheet formed from Ala115 to Ala120. Thus, structure A suggests that IAPP and PrP106-126 prefer to adopt random structures. The binding energy estimated using MM/PBSA showed relatively large contribution from van der Waals interactions, which confirms that hydrophobic effects are the main driving force.

From Figure 11, we observe that in structure B, D and F, IAPP adopts helical structure near the N-terminus. Structure B contains extended stretch of 3_{10} -helix from Thr4 to Val17. Structure D is the only complex adopting α -helix. However, Table 3 shows that it has a very low population of 1.3% with positive binding energy estimated by MM/PBSA mainly due to unfavorable electrostatic contribution. In addition, we have performed constant temperature molecular dynamics on structure A-F using Amber99SB-ILDN forcefield in order to test the robustness of our choice of forcefield and the effectiveness of dPCA in extracting important structures.¹⁶⁸ We carried out simple RMSD and hydrogen bond analysis (Figure S7, Table S2 and S3 in supporting information) and found that except for structure D, all the other structures are stable. Thus structure D can be considered as an outlier caused by the choice of bin size in plotting the free energy diagram. The stability of structure A, B, C, E and F show that our free energy landscape is not an artifact of the overstabilization of protein-protein interaction reported recently by Petrov and Zagrovic.¹⁶⁹

Due to the limitations of dPCA mentioned above, some important structures may be overlooked. Thus, the conventional way of conformational clustering was employed to explore other statistically significant structures from the 315 K trajectories. GRO-MOS clustering algorithm was used with a cut-off of 0.4 nm.¹⁵⁰ Due to the limitation of the clustering algorithm, only 20,000 frames can be processed each time, whereas our 315 K trajectory contains 500,000 frames from 100 ns to 600 ns. As such, the trajectory was cut into 25 sub-trajectories, each containing 20,000 frames. The time spacing

between each of the 20,000 frame is 25 ps. All the 25 sub-trajectories were submitted for clustering analysis. It was found that the top clusters for all 25 sub-trajectories were identical, suggesting that the clustering is robust.

Unsurprisingly, we found overlap between dPCA and conventional clustering. Structures A, B, C, E and F from dPCA are also identified in the conformational clustering analysis, again showing the robustness of dPCA. Most of the remaining structures extracted were found to have weak binding and random structures, and thus are included in supporting information. The top nine clusters with population of more than 2% were extracted and their representative structures are shown in Figure S5 in supporting information. An interesting structure G was found through clustering based on the observation that both structure B and structure G contain β -sheets at the same location and are highly similar to each other, except for the 3_{10} -helix in structure B. As structure B has a larger population than structure G, therefore it is likely that structure B transform into or from structure G by helix deformation or formation respectively. Structure H contains PrP106-126 in the β -hairpin conformation interacting with the N-terminal α -helix of IAPP. These two structures G and H are shown in Figure 11.

3.3.4 β -sheets Formation

As inter-chain β -sheet is closely linked to cross-fibrillation, it is thus an major focus of our study. In this context, other than extracting highly populated clusters from the simulation through dPCA and conformational clustering, more information may be extracted by searching for lesser populated structures containing inter-chain β -sheet. As such, we carried out clustering analysis using only the frames containing inter-chain β -sheets (calculated using DSSP) extracted from the 315 K ensemble. GROMOS algorithm was used with cut-off of 0.4 nm. We selected the top 15 clusters containing structures with inter-chain β -sheets (Figure 12) from the clustering analysis. 7 of the 15 structures were already identified through dPCA and conformational clustering in

previous sections and as such they were discarded to avoid redundant analysis. The binding energies were again estimated using the MM/PBSA approach and the results are shown in Table 4.

Fig 12. (1-8) Snapshots containing inter-chain β -sheets structures. Figures were generated using PyMOL. (9) Secondary structures determined by DSSP for structures 1-8. The definition of colors and representation of structures are identical to those given in Figure 11.

	Population	ΔE_{elec}	ΔE_{vdW}	ΔG_{polar}	$\Delta G_{nonpolar}$	$\Delta G_{binding}$
1	1.6%	19.3 ± 1.0	-77.7 ± 0.6	51.8 ± 0.9	-9.0 ± 0.1	-15.7 ± 0.6
2	0.6%	-40.0 ± 1.0	-101.5 ± 0.6	106.8 ± 0.9	-11.2 ± 0.1	-45.9 ± 0.8
3	0.6%	-13.7 ± 1.0	-68.8 ± 0.7	79.5 ± 0.8	-8.7 ± 0.1	-11.8 ± 0.8
4	0.5%	-28.2 ± 1.4	-112.6 ± 0.7	101.8 ± 1.5	-11.9 ± 0.1	-50.9 ± 0.8
5	0.5%	-88.0 ± 0.6	-10.6 ± 1.6	94.4 ± 1.4	-9.9 ± 0.1	-14.2 ± 0.9
6	0.5%	-50.8 ± 1.4	-83.8 ± 0.6	106.7 ± 1.2	-9.4 ± 0.1	-37.4 ± 0.7
7	0.4%	-14.1 ± 0.9	-76.3 ± 0.7	79.2 ± 0.8	-8.5 ± 0.1	-19.7 ± 0.7
8	0.4%	6.8 ± 0.9	-65.6 ± 0.4	49.6 ± 1.0	-6.8 ± 0.1	-15.9 ± 0.5

Table 4. Binding free energy estimated using MM/PBSA approach with the individual components shown. The energies are rounded to one decimal place and shown in kilocalories per mol. The binding energy is decomposed into individual component by g_mmpbsa where ΔE_{elec} is the molecular mechanics (MM) electrostatic energy, ΔE_{vdW} is the MM van der Waals energy, ΔG_{polar} is the polar solvation energy and $\Delta G_{nonpolar}$ is the non-polar solvation energy.

As the structures in Figure 12 are ignored by dPCA and conformational clustering, they are expected to be rarely populated. Nevertheless, meaningful observations can be made. The results show that the simulation sampled an extremely wide variety of structures forming inter-chain β -sheets, which may provide hints on the aggregation pathway of IAPP and PrP106-126. Structure 1 for example, shows an interesting β -barrel-like structure held together by short anti-parallel β -sheets.

3.4 Discussion and Conclusion

In this study, residues Gln10, Arg11 and Leu12 of IAPP are of particular interest to us. This helical region has been shown to be important in membrane binding.^{170–172} These three residues have relatively high tendency to form α -helix (Figure 8) during the simulation. Nevertheless, they are observed to adopt β -sheet structure and are involved in inter-chain β -sheet formation with residues His111, Met112, Ala113 of PrP106-126 (See Structure E from dPCA). Such conversion between α -helix and β -sheet secondary structures provides important insights into the dimerization process. In addition to that, the possibility of forming inter-chain β -sheet as in structure E is a significant result because it strongly suggests that heterogeneous fibril species can exist between IAPP and PrP106-126, an important question we set out to answer in this study. The results of this work provide a strong basis for more computationally intensive simulations to be performed on higher order oligomers.

Structure E, which has thus far been the more interesting structure due to its relatively extensive formation of inter-chain β -sheets, warrants extra attention in our analysis. It is noteworthy that structure E was found to have binding energy similar to that of an IAPP dimer in another study.¹⁷³ Analysis on the binding energies estimated using MM/PBSA generally shows that the IAPP-PrP106-126 complex prefers to form β -sheets based on the energetic perspective. Specifically, Structure E is most representative of the finding as it has the strongest binding energy among all the structures identified in our study. In conjunction with the the results discussed above in which α -helix can convert to β -sheet, our study has also shed light into how helix destabilization is a necessary component in the dimerization process leading to cross-fibrillation. To understand this, we seek to obtain mechanistic hints on how the process can happen through contact analysis. The results from contact analysis in Figure 9 show that PrP106-126 tends to contact with IAPP through the N-terminal region. High amount of contacts were found to involve aromatic residue Phe37 of IAPP. This finding was

further emphasized by the inter-residues distance map in which three out of ten pairs of residues with closest average distance involve Phe37 of IAPP.

In addition, two contact hotspots on PrP106-126 and three on IAPP for interactions between the two peptides were identified in Section 3.3.2. As such, we hypothesize the formation of inter-chain β -sheet structure process as such: in the earlier stage, the N-terminus of PrP106-126 approaches IAPP. The contact causes the α -helix of IAPP to destabilize, subsequently allowing the process of contact formation through various regions such as the hotspots identified in Section 3.3.2. Conversion of secondary structures involving Gln10, Arg11 and Leu12 of IAPP provides the platform for the formation of inter-chain β -sheet to occur. The entire process suggests the requirement of high degree of precision in the geometrical orientation of both proteins, rather than just high local concentration as was initially proposed. Indeed, the global minimum in our simulation, structure A from Section 3.3.3 was found to be random, indicating that in the absence of membrane, PrP106-126 and IAPP do not favor β -sheet formation.

Incidentally, the fact that PrP106-126 and IAPP do not share sequence similarity provides an opportunity to investigate the significance of aggregation between heterogeneous protein species. We expect heterogeneous aggregation to provide a higher diversity of amino acids for peptide-peptide interaction, leading to higher structural polymorphism in a protein complex. Various experimental and simulation studies have been done on homogeneous aggregation of both PrP106-126 and IAPP, notably the computational study of oligomerization of PrP106-126 by Ning et al. and Grabenauer et. al, in addition to the simulation work by Laghaei et al. on dimerization of IAPP.^{91,110,111} In the paper by Ning et al., monomeric and dimeric PrP106-126 was not well characterized by any secondary structures and had almost no tendency to form β -sheet, although β -hairpin structure was observed in a trimeric system. On the other hand, the study of IAPP dimer by Laghaei et al. found the N-terminal region to favor helical conformation, but contains almost no β -sheet in the region starting from N-terminus to Val17. In our study of heterogeneous aggregation however, both

PrP106-126 and IAPP show moderate tendency to form β -sheet throughout the full sequence. Both peptides, however, have sampled large amount of random coils and turns. While detailed structural comparison between literatures and our study should be taken carefully due to the potential bias introduced through the use of different force fields,¹⁷⁴⁻¹⁷⁶ it is clear that our simulation have indeed sampled a relatively more diverse phase space.

The slight dip in β -sheet propensity for PrP106-126 from residue 114-118 indicates that PrP106-126 can form β -hairpin. This is consistent with the finding of Ning et al. and Grabenauer et al., in which β -hairpin has been shown to be an important secondary structure in amyloid formation.^{110,111} Three representative β -hairpin structures is observed in Figures 11B, 11G and 11H. β -hairpin can also be observed in structure 5 in Figure 12. The two β -strands of β -hairpins in aforementioned structures are centered around the palindromic region AGAAAAGA from residue 113-120, therefore in agreement with experiments and simulations. This hydrophobic region was shown to be important in mediating the secondary structures of PrP106-126.¹⁶⁴ Structure H in particular illustrates a well-structured hairpin interacting with the N-terminal α -helix of IAPP.

Let us now take the perspective from IAPP. We found that our simulation has equal propensity to sample helix at the N-terminal region from Thr4 to Leu16 and β -sheet from Gln10-Leu16 for the IAPP protein. These include the three residues Gln10, Arg11 and Leu12 involved in the inter-chain β -sheet mentioned above. One particular region worth investigating is the SNNFGAIL (residue S20-L27) region of IAPP, which has been shown to be amyloidogenic.^{82,165} In our simulation, SNNFGAIL region is found to favor β -sheet structure (Figure 8). It often involves in contacts with PrP106-126 and in fact belongs to two of the six contact hotspots identified in Figure 9. Thus the simulation suggests that SNNFGAIL is also important in heterogeneous aggregation and not just homogeneous aggregation of IAPP.

Structurally, Buchanan et al. found the importance of the FGAIL region in IAPP

to form transient β -sheet in the formation of IAPP aggregates.⁸² Young et al. detailed the different mechanism of SNNFGAIL in mediating IAPP aggregation.¹⁶⁵ In Figure 11, we observe inter-chain β -sheet formation in residue Ile26 and Leu27 for structure B and residue Asn21 to Phe23 for structure E. Our analysis of transient β -sheet formation identified structures 3, 5 and 6 (figure 12) where β -sheets are observed to form at the SNNFGAIL region. Interestingly, in structure 3, FGAIL forms parallel inter-chain β -sheet with Lys110 to Ala115 of PrP106-126. Parallel β -sheets may be important in the seeding process as the protofibril of A β , for example, is usually in the form of parallel β -sheet.⁴ A study by Dupuis et al. found β -hairpin to be a main structure in dimer assembly of IAPP.¹⁷⁷ While our simulation does not observe such extensive β -hairpin, short hairpin can be found in structures B and G at the C-termini, and in structure E between residue Asn21-Phe23 and residue Ser29-Asn31. Finally, two of the major contact hotspots of IAPP centered around residues 8-18 and 22-28 uncovered by us have been determined experimentally to be important for both self association of IAPP and hetero-association with A β .¹⁷⁸

3.5 Supplementary Information

3.5.1 Comparison of dPCA Energy Landscape at Three Lowest Temperatures

We can see that for 317 K and 319 K the proteins tend to sample more of the non-minima structures (More broad, dark red regions). The general shape (largely determined by the eigenvectors of PC1 and PC2) and the deepest minima, however, remain the same. Thus it indicates that at higher temperatures there's more smearing of the minima, which may introduce unnecessary noise and bias if they were to be incorporated into the analysis together with the lowest temperature.

Fig S1. Comparison of the dPCA energy landscape for the three lowest temperatures ensembles.

3.5.2 The Evolution of into Structure E

In attempt to reveal the mechanism of the association of PrP106-126 and IAPP into the β -sheet structures in structure E from dPCA, we checked the trajectories containing continuous atomic coordinates instead of constant temperatures trajectories. This was

done by tracking the replica instead of the temperature ensembles. The continuous atomic coordinates trajectory of REMD has been used to yield important insight in the structural evolution despite that the temperature of such trajectory vary throughout the simulation, for example to establish transition map.¹⁷⁹ We analyzed a replica that evolved into the structure E. Readers are reminded to exercise caution when using the analysis following because strictly speaking, the trajectory is not meaningful thermodynamically due to the fluctuating temperatures.

A few important stages are shown in Figure S2. In Figure S2(a), the first contact of the proteins was a hydrogen bond formed between Lys5 of PrP106-126 and Cys2 of IAPP. PrP106-126 can be seen to form hairpin within itself. Then, the two proteins collapse into a compact structure as observed in Figure S2(b). Hydrogen bond was formed between Ala13 of PrP106-126 with Leu12 of IAPP. The region close to the C-terminal of IAPP was observed to form helical structures intermittently. Then in Figure S2(c), Met4, Lys5, His6, Met7 and Ala8 of PrP106-126 formed α -helix. Gly9 and Ala10 of PrP106-126 forms transient inter-chain β -sheet with Gln10 and Arg11 (colored in yellow in figure S2(c)) of IAPP. After that, in Figure S2(d), the contact between Ala13 of prion with Leu12 of IAPP started to break apart, allowing His6 of PrP106-126 to come closer to Leu12 of IAPP. At this stage, Asn14 to His18 of IAPP can be seen to form α -helix (colored in orange in Figure S2(d)), limiting the flexibility of the structures. However, as the simulation continues, the α -helix destabilizes, allowing the regions come close to involve in inter-chain β -sheets formation. The intra-chain β -sheets for IAPP was also formed at this stage. Lastly, the backbone of PrP106-126 and IAPP slides alongside each other, finally aligning to form the inter-chain β -sheets shown in structure E. (The numeration of PrP residues should be changed to the new system as in the manuscript.)

Fig S2. Snapshots of the different stages described in the structural evolution of PrP106-126 and IAPP complex. Yellow ball indicates N-terminal. Blue is used for PrP106-126 chain, while red is used for IAPP chain. t denotes the time of the frame, T indicates the temperature bath the replica was in at the particular time. Figures were generated with PyMOL.¹

Analysis of the structural evolution pathway using continuous trajectory hinted that at one stage of the simulation, a helix was hindering the formation of β -sheets. The high temperature heat bath enabled by REMD was able to unfold the helix. The α -helix to β -sheets transition was consistent with a previous study by Goh et al. on proteins folding using CSAW model.¹⁸⁰ Other than that, during stage (a) (Figure S2) we found that the first contact between the two peptides involved Cys2 of IAPP. This

is important as the disulfide bridge between Cys2 and Cys7 has been shown to be important in stabilizing the helix at the N-terminal of IAPP.¹⁸¹

3.5.3 DSSP Plot for 315 K Ensemble

(a)

Fig S3. DSSP plot for ensemble 0 showing the diverse sampling throughout the REMD simulation.

3.5.4 Dynamical Reweighting using MBAR

As we can see, structure A-F determined by using only the lowest energy ensemble (refer to main manuscript) can also be found here and the general feature of the PMF is similar to the original plot. The similarity is to be expected for a well converged REMD simulation. Chodera et. al mentioned that for the reweighting to work well with the simulation, the exchange time should correspond to the correlation time of the system.¹⁶⁷ With no prior knowledge of the autocorrelation structure of the system, it is not easy to fulfill the criteria, especially for a large and complicated system in our case (55 replicas) where trial and error is not a feasible strategy. Second issue for our system is that we have a large number of frames, thus if we assume all frames

to be uncorrelated, the amount of memory requires for the computation (55 replicas * 500000 frames) far exceed the capability of the cluster we have. Nevertheless, we attempt reweighting using PyMBAR by Chodera et. al by assuming the correlation time to be 100 frames in order to reduce the amount of frames to reasonable number. Skipping frames with the arbitrary choice of 100, however, results in some of the high energy region to be lost as evident by comparing figure S4 with that of the plot in the main manuscript. Artificial minima may also arise for the same reason. Lastly, for the free energy value, we have mentioned that due to the issue of certain minima containing multiple structures, the energy value should be interpreted carefully. We instead calculated the population of structures by counting with RMSD in the original manuscript. The same issue exists too with reweighting. Thus the advantage of dynamical reweighting in potentially giving more accurate free energy value is lost here.

(a)

Fig S4. 2D potential of mean force plot using dynamical reweighting on all 55 ensembles from 100 ns to 500 ns.

3.5.5 Structures from Conformational Clustering

Fig S5. Top 9 structures from conformation clustering method described in main manuscript. The numbering of the structures corresponds to the population number in descending order.

Table S1. Binding free energy estimated using MM/PBSA approach. Data for structures a, b, c, d and h were taken from table 2. The binding energy was decomposed into individual component by g_mmpbsa where ΔE_{elec} is the molecular mechanics (MM) electrostatic energy, ΔE_{vdW} is the MM van der Wals energy, ΔG_{polar} is the polar solvation energy and $\Delta G_{nonpolar}$ is the non-polar solvation energy.

	Population	ΔE_{elec}	ΔE_{vdW}	ΔG_{polar}	$\Delta G_{nonpolar}$	$\Delta E_{binding}$
a	8.50%	-37.2 ± 0.4	-89.7 ± 0.2	109.7 ± 0.4	-10.8 ± 0.0	-28.0 ± 0.3
b	6.00%	14.3 ± 1.2	-79.5 ± 0.3	51.5 ± 0.5	-8.9 ± 0.0	-22.6 ± 0.3
c	5.40%	9.2 ± 0.9	-109.4 ± 0.6	72.5 ± 0.8	-11.1 ± 0.1	-38.7 ± 0.6
d	3.90%	-18.1 ± 1.0	-117.3 ± 0.9	89.5 ± 1.0	-12.1 ± 0.1	-57.9 ± 0.9
e	3.60%	1.9 ± 0.8	-75.5 ± 0.9	73.4 ± 0.9	-8.6 ± 0.0	-8.8 ± 0.6
f	3.40%	43.8 ± 1.0	-56.0 ± 0.6	29.7 ± 0.9	-6.1 ± 0.1	11.4 ± 0.7
g	3.00%	-4.4 ± 0.7	-68.0 ± 0.6	60.4 ± 0.7	-7.0 ± 0.1	-19.1 ± 0.5
h	2.40%	-17.3 ± 0.8	-123.2 ± 0.8	101.8 ± 0.8	-12.8 ± 0.1	-51.5 ± 0.7
i	2.20%	6.3 ± 0.7	-61.0 ± 0.7	47.4 ± 0.9	-7.0 ± 0.1	-14.3 ± 0.6

Fig S6. Percentage population of all replicas visiting the middle (30th ensemble of the 55 total ensemble) ensemble simulated at 382 K. This shows that replicas from both left and right of the ensemble visited the middle ensemble with a random pattern in general.

3.5.6 Constant Temperature Molecular Dynamics with Amber99SB-ILDN

Table S2. Average hydrogen bond number between IAPP and PrP106-126 for structure simulated with OPLS-AA forcefield. The value was calculated using g_hbond utility. Note that structures A-F were obtained from our REMD simulation through dPCA analysis.

A	B	C	D	E	F
6.933793	7.465882	15.317659	4.691641	11.611025	13.560795
1.719192	1.767776	1.894651	1.374004	1.924183	2.329793

Table S3. Average hydrogen bond number between IAPP and PrP106-126 for structure simulated with Amber99SB-ILDN forcefield. The value was calculated using g_hbond utility. Note that structures A-F were obtained from our REMD simulation through dPCA analysis.

A	B	C	D	E	F
10.763905	10.893082	13.849983	5.49562	13.180816	13.171297
2.580075	3.071413	3.034161	1.407775	2.147611	1.74182

(a)

Fig S7. Root-mean-square deviation of distances between backbone atoms for structure A-F obtained from dPCA analysis and then simulated using Amber99SB-ILDN using constant temperature molecular dynamics.

4 Towards a Mechanistic Understanding of Peptides Gelation using Sequences from Sucker Ring Teeth Proteins

4.1 Introduction

In collaboration with Ali et al., we recently discovered a short 8 amino acids sequence GV8 (Ace-GLYGGYGV-NH₂) that can form hydrogel by mixing with water. The sequence is part of the sucker ring teeth (SRT) protein. Study of this protein is inspired by the predatory efficiency of squid and cuttlefish. It was discovered that the sucker ring teeth consists of many modular peptide sequences that are able to form β -sheets which in turn gives rise to the strong mechanical properties of sucker ring teeth.^{182,183} Our study was further inspired by the fact that another highly similar sequence GL8 (Ace-GLYGGYGL-NH₂) aggregated into dense agglomerates instead of forming gel. Only a single amino acid located at the C-termini of these two peptides differ, with the valine of GV8 being substituted by leucine in GL8.

Circular dichroism study shows that both peptides possess signatures that are representative of β -sheets structures. Leucine is more hydrophobic than valine, and hydrophobicity has been found to affect aggregation kinetics of short peptides.^{184,185} Bowerman et al. studied peptide amphiphiles with sequence XKXE where X is substituted with phenylalanine, alanine, valine, leucine and cyclohexylalanine.¹⁸⁴ Peptide amphiphiles have been known to form β -sheets hydrogels and while aromatic residues were traditionally recognized as a key factor in the hydrogelation process, it was found that by using amino acids with different hydrophobicity, both the morphologies and secondary structures could be regulated. In Bowerman et al.'s study, both peptides substituted with valine and leucine were able to form β -sheets and fibrils under transmission electron microscope (TEM). Peptides with leucine, however, contains globular

amorphous aggregates. The results are thus similar to our peptides to a certain extent.

In this ongoing project, we aim to elucidate how a single amino acid may ultimately affect the self-assembly of peptides GV8 and GL8 via computational simulation. The supramolecular structures of peptide gel require modelling of enormous amount of peptides and long sampling time, both of which are not within our computational capacity. Using coarse-grained modelling, we first look at the self-assembly process of a large system of 200 peptides. The peptides were allowed to form stable supramolecular assembly during the simulation. Then, using atomistic modelling and enhanced sampling methods, we investigated the early stage of self-assembly of peptide GV8 and GL8 by simulating both peptides in tetramer configuration.

4.2 Methods

4.2.1 Martini Coarse-grained Model

Martini model was used to carry out coarse-grained molecular dynamics simulation of peptides.^{131,132} Martini forcefield aims to model biomolecules by using a single bead to approximate interactions of a group of atoms. On average, four atoms are represented by one bead. This allows faster sampling of conformations at the expense of fine atomistic details. In essence, coarse-grained model has the effect of smoothing out rough free energy landscape. We have also employed polarizable water molecules developed for Martini model to better model the partial charges of water molecules.¹⁸⁶ All simulations were carried out by inserting 200 peptides into a 15 nm x 15 nm x 15 nm box with $\approx$ 24k polarizable Martini water molecules. Martini model requires the secondary structures of the peptides to be set prior to the simulation. Our collaborator has provided us preliminary results from circular dichroism study of the peptides (Unpublished), of which both peptides are estimated to contain approximately 40% β -sheets, 15% α -helix. The remaining 45% was set to be random structures. Temperature was set at 303 K. This coarse-grained model also allows an integration time step of 25 fs.

Simulations were carried out by first minimizing the potential energy of the system using steepest descent algorithm in order to avoid steric clashes. Then the system was equilibrated to desired temperature for 1250 ps. Finally, production run was carried out for 750 ns. As the Martini model is known to accelerate sampling by a factor of four,^{187,188} thus the effective simulation time for our system is approximately 3 μ s. Three runs were repeated for each simulation to ensure the results are reproducible and at the same time allows us to have 9 μ s of effective simulation time. The systems correspond to peptides' concentration of $\approx$ 400 mM which is about an order of magnitude higher than the experimental condition. The higher concentration was used for two reasons. Firstly, within limited simulation time it will accelerate the aggregation

process. Secondly, for a similar box size, 20 mM concentration will result in just two peptides in a box, which in turn does not produce meaningful insights into the gelation process. On the other hand, it will require an extremely large box to be able to accommodate 200 peptides at 20 mM concentration, which is beyond our computational capabilities. We have also simulated both systems in smaller box at 10 nm x 10 nm x 10 nm ($\approx$ 2000 mM) using similar protocol to investigate the concentration effect of peptide.

4.2.2 Bias Exchange Well-tempered Metadynamics (WT-BEMD)

In any proteins simulation, one of the most important issues is to be able to sample the conformational phase space as exhaustive as possible. A classical molecular dynamics simulation has difficulties crossing high energy barriers within the time scale limited by current computational capabilities (Chapter 2). Further compounding the issue is the enormous amount of possible conformations even for a small peptides.¹⁸⁹ Over the past decades, several methods were developed to enhance sampling speed of molecular dynamics.^{190,191} One of the most well-established methods available currently is metadynamics.⁶⁷

The fundamental working principles of metadynamics rely on the concept of collective variables, or sometimes also known as reaction coordinates. A collective variable attempts to describe a system in its most important degree of freedom, e.g. dihedral angles of a specific residues, amount of secondary structures, number of hydrogen bonds etc. These information has to be known a-priori to any simulation. A potential mean force (PMF) diagram can then be plotted to describe the system of interest in its collective variable and the corresponding free energy. Metadynamics make use of this information by adding bias potential to the existing potential V with respect to the collective variable:

$$V_G(S(x), t) = \omega \sum_{t'=\tau_G, 2\tau_G, \dots}^{t' < t} \exp\left(-\frac{(S(x) - s(t'))^2}{2\delta s^2}\right) \quad (4.1)$$

$$V_G^{WT}(S(x), t) = \omega \exp\left[-\frac{V_G(S(x), t)}{k_\beta \Delta T}\right] \sum_{t'=\tau_G, 2\tau_G, \dots}^{t' < t} \exp\left(-\frac{(S(x) - s(t'))^2}{2\delta s^2}\right) \quad (4.2)$$

where $S(x)$ is the value of collective variable, ω is the gaussian height, τ_G is the gaussian deposition stride, $s(t')$ is the value of collective variable at time t and lastly δs is the gaussian width. Equation 4.2 is the so called well-tempered metadynamics (WT-MetaD), where the potential V_G is scaled by a factor depending on the history of gaussians already deposited and a bias factor ΔT with the unit of temperature. By scaling the gaussian to nearly zero as time proceeds, well-tempered metadynamics ensures the free energy landscape to converge smoothly without high statistical errors compared to the normal variant of metadynamics (equation 4.1).^{192–194}

In this study, we employed 7 generic CVs commonly used for proteins folding in our simulations. The collective variables are: α -RMSD, anti-parallel β -RMSD, parallel β -RMSD, coordination numbers between hydrophobic side chain atoms, backbone hydrogen bonds, coordination numbers between α -carbon and radius of gyration of α -carbon. The secondary structures collective variables are defined as:

$$s = \sum_i \frac{1 - \left(\frac{r_i - d_0}{r_0}\right)^n}{1 - \left(\frac{r_i - d_0}{r_0}\right)^m} \quad (4.3)$$

where r_i is the root mean square deviation of any segment i of a protein with respect to an ideal secondary structure segment of equal length (3 + 3 contiguous residues for β -RMSD and 6 residues for α -RMSD). r_0, d_0, n and m are tuning parameters for the sum function. Following the original paper, the parameters are set as 0.1 nm, 0.0 nm, 8 and 12 respectively.¹⁹⁵ Equation 4.3 can be used to calculate coordinate numbers as well by replacing r_i with r_{ij} and choosing r_0 to be the reference distance for the

contacts interested. For hydrophobic side chain atoms contact, backbone hydrogen bonds and α -carbon contacts, r_0 is set as 0.5, 0.25 and 0.65 respectively similar to a paper by Laio et. al.¹⁹⁶ Radius of gyration $\mathbf{R}_G$ is defined as:

$$\mathbf{R}_G = \left(\frac{\sum_i^n m_i |r_i - r_{COM}|^2}{\sum_i^n m_i} \right)^{\frac{1}{2}} \quad (4.4)$$

where m_i is the mass of i^{th} atom, r_i is the coordinate of i^{th} atom and r_{COM} is the center of mass of the atoms used in the calculation.

These collective variables have been used successfully in folding proteins within reasonable timescale.^{195–197} 8 replicas (including a neutral replica without bias) were simulated simultaneously for 1.2 μ s resulting in effective simulation time of 9.6 μ s. Bias factors for all collective variables were set to be 10. Gaussians were deposited every 2 ps. Gaussian heights were set to be 1 kJ. Gaussian widths were set to be 0.20, 0.30, 0.30, 0.30, 2.00, 5.00, 0.10 for each of the seven CVs respectively. PLUMED 2.1 was used in conjunction with GROMACS 4.6.7 for all enhanced sampling simulations.^{150,198,199} For molecular dynamics, we used AMBER99SB-ILDN forcefield.²⁰⁰ A short molecular dynamics simulations were carried out to obtain an initial structure of the complex formed by the tetramer. Then, it was put into a dodecahedron water box with a minimum distance of 1 nm from each side of the tetramer. Periodic boundary conditions were applied on the three dimensions of x, y and z. Ions were added to maintain an ionic concentration of 0.15 M. GROMACS 4.6.7 was used to perform the simulation.¹⁵⁰ LINCS protocol was used to constrain bonds involving the hydrogen atoms to allow integration time step of 2 fs.¹⁵¹ Particle Mesh Ewald with cut-off of 1.0 nm was used.¹⁵² 1.0 nm cut-off was used for vdW interactions. Coordinates were saved every 2 ps.

4.3 Results

4.3.1 Coarse-grained Simulations using MARTINI Model

We first carried out simulations of large number of both GV8 and GL8 peptides using Martini model, a coarse-grained molecular dynamics method described in section 4.2.1. This allowed us to have an insight into the higher order structural features of the peptides. This was inspired by Frederix et al. study of hydrogel using Martini model in which it was found that the model was able to predict self-assembly of tripeptides into hydrogel.¹³⁰

First, to check if the general feature of the simulation can be repeated, we calculated solvent accessible surface area (SASA) of the final complex formed by peptides GV8 and GL8. Solvent accessible surface area is defined as the surface area of residues that is in contact with the solvent.²⁰¹ Figure 13 shows the Solvent Accessible Surface Area (SASA) of the simulated peptides. As we can see the peptides aggregated quickly and their SASA converge to a stable value after a short period of time. All three repeated simulations for each of peptide GV8 and GL8 converged to approximately the same value, hence showing the reproducibility of the self-assembly dynamics.

In order to look at how the last residue of the different peptides arrange in the aggregated complex, we calculated the radial distribution function using the backbone atoms of the last residue throughout the simulation. The radial distribution functions $g(r)$ of all the systems are plotted in Figure 14. The value of the radial distribution function indicates the probability of a particle within the sphere with a particular radius compared to that of an ideal gas. Both peptide GV8 and GL8 have similar radial distribution function, with the maximum peaks located at ≈ 0.5 nm. Figure 15 shows the snapshots of the simulated peptides at 750 ns. We observed a few interesting nanostructures formed by the peptides at the end of the simulation. Specifically, Both GV8 and GL8 aggregate into thick rod or spherical structures where the peptides do not have special preference for the backbone orientation.

Next, the shapes of the nanostructures can be quantified by calculating the radius of gyration in x, y and z component. Radius of gyration is defined as in equation 4.4. Radius of gyration has been traditionally used to quantify the compactness of protein. However, it is also related to the moment of inertia as such:

$$\mathbf{I} = \sqrt{\frac{\mathbf{R}_G}{m}} \quad (4.5)$$

where m is the mass of atoms. by looking at its x, y and z component, the simple shape of a protein (in our case, aggregated peptides) can be determined. For example, we are interested in nanofibres formed by the peptides which will result in $I_x \ll I_y \approx I_z$. The values of radius of gyrations and moments of inertia for the last frame of each of the siulations are tabulated in table 5. We can see that while the simulations resulted in similar radial distribution function and solvent accessible surface area, the nanostructures may adopt different shapes. Run two of peptide GV8 resulted in a twisted nanofiber, whereas run one and three resulted in flat disk-like nanostructures. Comparatively, all three runs of peptide GL8 resulted in $\frac{I_z}{I_y}$ values closer to one (i.e. the first criteria $R_{Gz} \approx R_{Gy}$). The second criteria of $\frac{I_z}{I_x}$ indicates how elongated the fiber structures are if the first criteria is satisfied. Thus we can see that In general, peptide GL8 adopted a more fibril like structure (average value of $\frac{I_z}{I_y}$: 1.046 ± 0.026) and while GV8 (average value of $\frac{I_z}{I_x}$: 1.237 ± 0.076) was able to form a twisted nanofiber, two other runs resulted in disk-like nanostructure. Our collaborator used scanning electron microscope (SEM) to examine both peptides and found that GV8 adopted morphology of sheets of assembled peptides while GL8 was strongly fibrillar.

Next, the peptides were simulated in an artificially high concentration of ≈ 2000 mM by placing 200 peptides in a 10 nm x 10 nm x 10 nm box. This allows the peptide to interact via periodic boundary condition. At high concentration, interesting

structures resembling network of rod-like structures emerge when periodic boundary condition is taken into consideration. Waters can be seen to be enclosed within the peptides network (figure 15). However, the same structures emerge for both peptide GV8 and GL8 in contrast to experiment in which only peptide GV8 was able to form gel, highlighting the shortcoming of coarse-grained modelling.

Fig 13. Solvent Accessible Surface Area (SASA) for the different peptide systems simulated. Each color represents a repeat of the simulation for the same peptide system. (a) GV8. (b) GL8. (c) GV8 at 2000 mM. (d) GL8 at 2000 mM.

Fig 14. Radial distribution function $g(r)$ for (a) GV8 (b) GL8 (c) GV8 at 2000 mM and (d) GL8 at 2000 mM concentration.

Table 5. Table shows moments of inertia for the final snapshots of simulated peptides GV8 and GL8 at 400 mM concentration. Units for radii of gyration are in nm . Units for moments of inertia are in $a.m.u.nm^2$.

		R_G	I	I_x	I_y	I_z	$\frac{I_z}{I_y}$	$\frac{I_z}{I_x}$
GV8	Run1	3.223	841766	362332	453877	609328	1.342	1.682
	Run2	3.610	1096090	301489	687043	799062	1.163	2.650
	Run3	3.174	824953	301015	490554	591011	1.205	1.963
GL8	Run1	3.320	913905	283139	606910	621863	1.025	2.196
	Run2	3.435	998548	253719	672690	692974	1.030	2.731
	Run3	3.199	834054	310338	525377	568608	1.082	1.832

Fig 15. Representative structures for (a)(c) Run one to run three of peptide GV8 in 400 mM concentration and (d)(f) Run one to three of peptide GL8 in 400 mM concentration. Water molecules are not shown here as they are not important to the visualization. (g) Scanning electron microscope (SEM) images of both peptides GV8 and GL8 (Image provided by Shu Hui, Hiew and Ali Miserez).

Fig 16. Representative structures for the final structures obtained in simulations with 2000 mM concentration of (a) peptides GV8 and (b) peptides GL8. Water molecules are colored in cyan. Periodic boundary conditions were shown by drawing the neighbouring boxes in x, y and z direction.

4.3.2 Convergence of Metadynamics Simulation

With the insights obtained using Martini modelling, we then used molecular dynamics as discussed in section 4.2.2 to gain atomistic insights into inter-peptides interactions. Eight replicas were simulated for 1.2 μ s each. We used the last 400 ns for each replica in our subsequent analysis. This was chosen based on the fact that there is no significant gaussian hills added after 800 ns of simulation, i.e. there is no new region being explored in the phase space described by the seven CVs (Figure S8 and S9) except for hydrophobic side chain contacts. Both GV8 and GL8 were sampling very high hydrophobic side chain contact towards the end of the simulation. The sampling was however very short (Figure S8d and S9d), indicating that they were unstable. In fact, the free energy plot in the following discussion was not affected by the sampling of these structures.

We adapted the tool METAGUI to analyse our simulation.²⁰² The convergence of the simulation was checked by plotting the free energy curve for the first and second half of the last 400 ns used for our analysis. The potential mean force (PMF) curves are plotted for each CV and are shown in figure 17 and figure 18. We can see that the

free energy curves in the first and second half of the last 400 ns of the simulation did not deviate significantly from each other, therefore our simulations were able to achieve reasonable convergence within the phase space sampled. The green curve represents the "connected" region where it indicates that in the connected regions, the free energy sampled in the first and second half of the simulations does not deviate from each other for more than $6 k_{\beta}T$ (≈ 12 kJ). In GV8's simulation, we note that structures with high α -RMSD content (> 6 in figure 17(a)) were not sampled in the last 400 ns as shown in figure 17a, indicating some inefficiency in the metadynamics simulation. Note that the regions that were not connected (not joined by green curve) will be discarded by METAGUI in the analysis as they were deemed to have unsatisfactory convergence. For GV8, this includes the structures with high α -RMSD content mentioned above, which would not change our discussion as we are most interested in formation of β -sheets.

Fig 17. Free energy for the 7 CVs used in the simulation for peptide GV8. Magenta line represents free energy sampled during the first 200 ns of the last 400 ns of the simulations, while blue line represents free energy sampled during the second 200 ns of the last 400 ns of the simulations. Green line represents the average value of the magenta and blue line, with the additional condition that it will only be drawn if the deviation between magenta and blue line is lower than $6 k_B T$ (a) AlphaRMSD. (b) Anti-parallel β -RMSD. (c) Parallel β -RMSD. (d) Side-chain contact of Leucine and Valine. (e) Backbone hydrogen bonds. (f) Contacts between α -carbon. (g) Radius of gyration.

Fig 18. Free energy for the 7 CVs used in the simulation for peptide GL8. Color of the lines are defined to be the same as that of figure 17. (a) AlphaRMSD. (b) Anti-parallel β -RMSD. (c) Parallel β -RMSD. (d) Side-chain contact of Leucine and Valine. (e) Backbone hydrogen bonds. (f) Contacts between α -carbon. (g) Radius of gyration.

4.3.3 Metadynamics: Difference in Free Energy Landscape

In addition to checking convergence, figure 17 and 18 showed the similarities and differences in the free energy landscape of β -sheet formation in both peptides. For anti-parallel β -sheet, the difference in PMF is relatively minor with GL8 having a slightly flatter well extending to ≈ 12 . Major difference can be found in PMF of parallel β -sheet formation, where GL8 has a local minimum located at ≈ 17.5 . Local minima are typically metastable states. Interestingly, no such local minimum exists in GV8. It should be noted however, that this local minima has a large free energy difference relative to the minima with lower parallel β -sheets in peptide GL8 and thus

should not contribute significantly to the ensemble average as we will calculate in the following section. Figure 17a and 18a showed that both peptides do not like to form α -helix. Lastly, Energy minima at ≈ 1 nm in the PMF of radius of gyration (figure 17g and 18g) is consistent with the hydrophobic nature of both peptides.

4.3.4 Metadynamics: Multi-dimensional Free Energy Clustering Analysis

Next, we attempted to look at structures with low free energy (stable) sampled during the simulation. Using METAGUI again, the trajectories was clustered using 7 CVs by assigning the conformations sampled in all trajectories into a 7-dimensional hypercubes. Then, using a weighted histogram analysis method (WHAM) approach described by Marinelli et. al.,²⁰³ free energies of the clusters were calculated taking into consideration the bias potential applied during the simulation. Note that as we are using well-tempered metadynamics in our simulation, the potentials accumulated are rescaled by a factor of $\frac{\gamma-1}{\gamma}$.²⁰⁴ We can then look at clusters from different minima of the 7-dimensional free energy landscape. As this is a simulation of intrinsically disordered peptides, the global minima of both systems are expected to be very broad and contains highly polymorphic structures. Specifically, since we are interested in how the peptides can adopt β -sheet structures, clusters containing β -sheets with low free energy are extracted and analysed. The representative structures from different region of energy minima are shown in figure 21. Interestingly, the structure with high anti-parallel β -sheets is a global minimum in the multi-dimensional free energy landscape calculated with all 7 CVs for peptide GL8, whereas peptide GV8 did not have a well-defined structures at the lowest free energy minimum. In particular, GL8 was observed to form a structure with highly ordered anti-parallel β -sheet (figure 21(C)).

Next, using the free energy assigned to different clusters, we were able to calculate

the ensemble average value of any quantity of interest using the following formula:

$$\langle O \rangle = \frac{\sum_{\alpha} O_{\alpha} e^{-F_{\alpha}/T}}{\sum_{\alpha} e^{-F_{\alpha}/T}} \quad (4.6)$$

$$\sigma^2(\langle O \rangle) = \frac{\sum_{\alpha} e^{-2F_{\alpha}/T} \left[\frac{(\langle O \rangle - O_{\alpha})^2}{T^2} \sigma^2(F_{\alpha}) + \sigma^2(O_{\alpha}) \right]}{\sum_{\beta} e^{-F_{\beta}/T}} \quad (4.7)$$

where O is the quantity of interest, O_{α} is the value of the observables for cluster α , F_{α} is the free energy calculated for cluster α and T is temperature. $\sigma^2(\langle O \rangle)$ is the error in the ensemble average, $\sigma^2(F_{\alpha})$ is the error in the free energy calculated for cluster α and $\sigma^2(O_{\alpha})$ is the variance of quantity O in cluster α .²⁰³

The ensemble averages for secondary structures, turn and coil explored during the simulation are shown in Table 4.3.4. Secondary structures for all trajectories were calculated using “do_dssp” utility according to the dictionary of protein secondary structure (DSSP).¹⁶¹ In general, we observe very similar properties for the two systems. This is again consistent with the experimental observations that the two peptides adopt similar secondary structures. The structures explored were mostly random and consist of only 5.90% and 7.74% β -sheet compared to $\approx 40\%$ in experiment estimated from CD spectra. This showed that fundamentally a tetramer simulation is unable to reproduce the supramolecular secondary structures content found in experiment.

In addition to that, the propensity of each residue in GV8 and GL8 to adopt β -sheet was evaluated using equation 4.6. Since the system consists of four identical monomer, the averages value of each amino acid was plotted in figure 19. As expected, the shapes of the plot indicated that the middle regions of the peptide sequences have higher tendency to form β -sheets. It can be observed that peptide GV8 and GL8 have similar β -sheets profile (within error bar) through most of the residues, except for TYR6. TYR6 is the second tyrosine residue in both peptide sequence. Here it can be seen that TYR6 has a significantly higher tendency ($\approx 10\%$ higher) to participate in β -sheets formation in peptide GL8.

The limitations of our simulations in using just a small tetramer may contribute

to the fact that the ensemble average values did not provide significant insights into the difference between peptide GV8 and GL8. Nevertheless, when we attempted to look at the clusters at different free energy cutoff, we were able to make compelling observations. First, we defined free energy cutoff at 1 kJ/mol , 2 kJ/mol , 4 kJ/mol , 8 kJ/mol , 16 kJ/mol , 32 kJ/mol and 64 kJ/mol . Then, we calculated the ensemble average of secondary structures for clusters with free energy lower than the cutoffs using equation 4.6. Results are shown in figure 20. It can be seen that at low free energy cutoff (more stable structures), peptide GL8 contains more β -sheets than peptide GV8 even though both peptides converged to the total ensemble average tabulated in table 4.3.4 at high energy cutoff as expected. Other than that, hydrogen bonds were calculated using “g_hbond” utilities and peptide GL8 was observed to form more hydrogen bonds at low free energy cut-off.

	α -Helix	β -Sheet	Coil	Turn	H-Bond Numbers
GV8	1.10%	5.90%	52.56%	8.01%	10.42
GL8	1.25%	7.72%	50.58%	9.28%	10.37

Table 6. Ensemble average values for the secondary structures explored during the simulation. Number indicates the average number of residues adopting the specific secondary structure. Errors are omitted here for clarity sake as they are negligible.

(a)

Fig 19. β -sheets propensity for every amino acid in the sequence. Error bars were calculated as the standard deviation of average values from four monomer for each amino acid.

Fig 20. Figure shows the ensemble average of (a) β -sheets structures (b) number of hydrogen bonds at clusters below different free energy cutoffs.

Fig 21. Representative structures from different energy minima. Arrows indicate the locations on the free energy landscape where the structures belong to.

4.4 Discussion

Bringing together the results from both atomistic and coarse-grained simulations, we are able to stipulate interesting insights into the gelation and aggregation process of peptide GV8 and GL8. From our experimental collaborators (Shu Hui et al.), data from circular dichroism (CD) (Unpublished) shows spectra similar to that of peptides with β -sheet structure, therefore highly ordered β -sheet structures in our simulation may be an important mechanism of which the peptides can further elongate into more extensive structure. Due to computational limitations, we were only able to simulate up to tetramer in the atomistic model and therefore unable to fully model the self-assembly mechanism. In this context, it is not surprising that the ensemble average values of important observables such as secondary structures and hydrogen bonds formation for both systems appear to deviate only slightly from each other, indicating that the two peptides are highly similar.

Nevertheless, it is interesting to note however that in the multi-dimensional free energy landscape, GL8 contains more ordered structures than that of GV8 in the lowest energy minima sampled during the simulation. This can be seen from figure 20. A tetramer represents a very early stage of self-assembly process for both peptides, thus it appears that while GL8 becomes highly ordered at low oligomer weight, this may not be favourable for the formation of peptide gels. In a study of silk fibroin peptide, the gelation process was proposed to be a three stage model.²⁰⁵ The early stage of gelation contains mostly hydrophobic interactions and backbone hydrogen bonds instead of ordered β -sheets. As the gelation process goes on, there will then be transition into higher order β -sheets complex. Our sequences are highly similar to the silk fibroin peptides sequence studied: GAGAGS, GAGAGY, GAGAGVGY, especially the presence of aromatic residue tyrosine which is a key factor in the formation of silk hydrogel. Both GV8 and GL8 contain two tyrosine residues at the second and sixth position in their respective sequences.^{33,206,207} Interestingly, in our simulation, we

found stronger participation of TYR6 (second tyrosine residue) in β -sheets formation for peptide GL8 as shown in figure 19. In a nutshell, the stronger hydrophobicity of peptide GL8 encourages more ordered β -sheets structure and affects the participation of tyrosine in the gelation process. Further simulation with high weight oligomer may shed lights into how these differences at low weight oligomer propagate.

Martini model has been used to study proteins/peptides aggregation and hydrogel formation.^{129,130,208–210} Most notably, Frederix et al. successfully predicted the aggregation of novel tripeptide sequences that can self-assemble into nanostructures.¹³⁰ We attempt to use similar protocol to elucidate the self-assembly mechanism of GV8 and GL8 peptide. In Martini model, both valine and leucine are modelled in similar fashion using a single hydrophobic bead as their side chain, with a small difference in the bond length between backbone atom and side chain. Our simulation results showed that at 400 mM concentration, peptide GL8 adopts nanostructures that are more fibrillar than peptide GV8, which may explain the morphology seen under scanning electron microscope. Interestingly, in a study by Bhak et al., it was found that α -synuclein that assembled into curly fibrillar structures was able to convert to amyloid hydrogel while those that assembled into straight fibrils could not.¹²³ In our simulation of GV8, one of the runs (run two) assembled into curly structure, which may play a role in the formation of hydrogel. However, we are unable to confirm the fibrils' morphology at 1 μm resolutions obtained by our collaborator.

Finally, we tried to simulate a much stronger concentrations of peptides as the formation of peptide networks at high concentration may provide a hint to how the peptides sequence may assemble into gel. This resulted in the network of nanofibres structures observed in figure 15. Nanofibres with high β -sheets content have been found to be one of the ways peptides gelate.^{36,211} However, the inherent limitation of MARTINI model does not seem to allow us to distinguish between the different hydrophobic amino acids in their formation of supramolecular complex at this high concentration. This was evident from the fact that both peptides GV8 and GL8 were

able to form similar structures. Nevertheless, we showed that the sequence motifs of both GV8 and GL8 have the tendency to form highly ordered network similar to silk fibroin proteins.

An issue commonly associated with the MARTINI model is its inability to model secondary structure transition. In this work, as we are attempting to model the final gel state of the peptides, it is reasonable to use MARTINI model. We do not expect MARTINI model to provide detailed dynamics of the aggregation process of the hydrogel. It should, however, provide reasonable projection into observable quantities such as the shapes of the final aggregated structure implied through the moments of inertia calculations. We acknowledge the limitations of MARTINI in modelling interactions caused by secondary structure changes and this was one of the main reasons for us to complement MARTINI simulations with atomistic simulations using metadynamics. NMR experiments are currently being carried out to provide us with more structural information that can be used to refine our simulations in the future.

4.5 Supplementary Figures

Fig S8. Figures show the potentials deposited throughout the 1.2 μ s simulation for peptide GV8.

Fig S9. Figures show the potentials deposited throughout the 1.2 μ s simulation for peptide GL8.

5 Binding Analysis of Profilin-Polyproline Complex

5.1 Introduction

In most eukaryotic cells, actin is known to be one of the most abundant proteins.²¹² It participates in a lot of vital body functions such as muscle contraction and supporting cell's cytoskeleton.²¹³ Nuclear actin is involved in various cellular functions around cells' nucleus such as the regulation of translation and transcription.²¹⁴ One highly important aspect of actins is its polymerization in forming F-actin, a filament cytoskeleton. F-actin structure has been solved experimentally and studied extensively.^{215–217} The early stage of actin polymerization involves actin nucleation which consists of a few actin monomers coming together to form a nucleus complex. Thereafter, more monomers approach and latch onto the nucleus before elongating into extensive actin filament.^{218,219} This polymerization process has been found to be regulated by the formin protein.^{219–221} In particular, the synergy between a protein called profilin (PRF) and the FH1 domain of formin protein is integral to the actin polymerization process. FH1 domain is located at the N-terminal region of formin protein and is an intrinsically disordered protein (IDP). While it does not possess a native structure, a consistent feature of FH1 domain is high amount of prolines in its primary sequence, which is known to be recognized by profilin as its ligand.^{222–225}

Structures of profilin in its apo configuration and in complex with polyprolines sequences have been studied extensively over the years.^{222,225–228} The folding motif of profilin typically consists of five-stranded antiparallel β -sheets and two α -helices. The α -helices are located at the N- and C- termini. From structures available in protein data bank (PDB), we can see that polyprolines bind close to the two α -helices. The binding orientation can be in two backbone orientations as shown by Mahoney et al..²²² (Figure 23(a)). In addition to that, profilin can bind to actin at another

binding site concurrently with polyprolines.²²⁹ Together, profilin acts as a regulator in the formin-mediated polymerization of actin.

In this project, we collaborated with Assistant Professor Miao Yansong from the school of Biological Sciences to better understand the binding mechanism between profilin (AtPRF) and a short polyprolines segment from FH1 domain (Manuscript in preparation).¹⁴³ Both profilin and polyprolines are from plant species *arabidopsis thaliana*. There are five profilin isomers expressed by *arabidopsis thaliana*.²³⁰ Miao et al. has studied the binding affinity and mechanism between AtPRF and polyprolines sequence **RVPPPPPPPPPLP**. The third profilin isomer AtPRF3 is particularly intriguing as it can potentially be expressed with a 37 residues N-terminal extension compared to a typical profilin sequence. In the experiment, AtPRF3 was synthesized in three versions: Full length with 37 residues N-terminal extension (AtPRF3-FL), moderately long N-terminal extension truncated at the 22nd residue (AtPRF3-N22) and lastly without N-terminal extension (AtPRF3-N37). It was discovered by Miao et al. experimentally that both AtPRF3-FL and AtPRF3-N22 exhibited stronger binding to polyprolines compared to the AtPRF3-N37.

Polyprolines is an important ligand to many proteins besides profilins. It refers to amino acid sequences that comprise of a high percentage of prolines joined together on a polypeptide. Among amino acids, proline is structurally unique due to its chemical composition. Unlike other amino acids, proline has a pyrrolidine ring joined to its α -carbon and α -amino group. This is illustrated in Figure 22. Due to its structure, proline has very rigid dihedral angles. In the most common *trans* isomer configuration, polyprolines form type II polyprolines helices (PP-II helices), with its angle ϕ and ψ restricted to ≈ -75 degrees ≈ 150 degrees respectively.²³¹ The structures of polyprolines have been studied using molecular dynamics by Moradi et al..²³² It was found that in water, the most probable structures for polyprolines were always PP-II helices. Their results were consistent with previous experimental results, thus lending credibility to the use of molecular dynamics to model polyprolines.^{233,234} As such, in

our model, polyprolines were initialized as PP-II helices.

Crystal structure of AtPRF3 in complex with polyprolines could not be obtained due to experimental difficulty. As discussed in Chapter 2, the structures of many proteins often cannot be obtained from experiments due to their insolubility and non-crystalline structures.³¹ However, atomistic models of such complex are needed to understand experimental results. Hence, we employed molecular dynamics to simulate AtPRF3-polyprolines complex in order to elucidate the binding mechanism. By surveying literature on the binding orientations and using homology models from available crystal structure of profilin from other species, we proposed potential binding modes and tested their stability using long molecular dynamics. Furthermore, the important interactions found were compared to literature to verify the validity of our models. We then used the MM/PBSA method to understand why AtPRF3-N22 bind stronger to polyprolines than AtPRF3-N37. By looking at the different components contributing to the binding free energy, we can also discover potentially crucial factors in the recognition mechanism of profilin.

Fig 22. Chemical composition of proline.

5.2 Methods

5.2.1 Homology Modelling and Molecular Dynamics

Homology modelling refers to constructing an atomistic model of a protein using its amino acid sequences based on highly similar and related proteins commonly referred to as a “template”.²³⁵ It has been successfully used to study WW and SH3 domains, both of which recognize polyprolines ligands.²³⁶ In our work, the homology model of the truncated version of AtPRF3: AtPRF3-N37 was created by SWISS-MODEL using PDB structure 1A0K as a template.^{228,237} 1A0K is the crystal structure of AtPRF1 from *arabidopsis thaliana* and hence we expect PRF3 to have similar fold. We chose to simulate AtPRF3-N22 instead of AtPRF3-FL because it contains a high number of charged residues close to the polyprolines binding site. In addition to that, simulating AtPRF3-FL would introduce a much higher computational complexity. This is due to the scaling complexity of $O(n^2)$ discussed in Chapter 2.

Then, polyproline peptide (polyprolines) was built using PyMOL as an extended chain with the sequence of **RVPPPPPPPPPLP** as per experiment.¹ To verify the quality of the built protein homology model, a quick 100 ns molecular dynamics simulation with explicit water was performed (details described below). The root mean square deviation of the distances (RMSD) between backbone atoms was calculated and found to be less than 0.15 nm (figure 23b), justifying the use of the homology model in the subsequent simulations. Next, we built the extended version of AtPRF3 by extending the N-termini of N37 with the sequence **QRRSRAKVKKKKKTN**. The extended amino acid sequences were modelled as an α -helix based on the prediction of PSIPRED webserver (figure S10 in supplementary information).^{238,239}

From the crystal structure of mammalian profilin 2a (PDB:2V8F),²²⁵ it can be seen that there are two potential binding grooves for polyprolines, in which it can align either parallel (L for lower position) or perpendicular (U for upper position) to the axis of N- or C- terminal helices. The initial binding pose of polyprolines was setup by

aligning the homology model to the crystal structure of 2V8F first, then subsequently aligning the built polyprolines to the polyprolines portion of the crystal structure's ligand. The alignment was done using PyMOL.¹ In addition to that, Mahoney et al. noted that polyprolines can also bind to profilin in two distinct orientations with the N-termini pointing in either in the forward or backward direction.²²² Therefore, we had also carried out simulations by reverse alignment of polyprolines from C- to N-termini to crystal structure's polyprolines.

Furthermore, the binding orientation is especially important for the polyprolines used in our study due to the positively charged arginine (R) residue located at the N-termini. This will be discussed in details in the later sections. The reverse alignment will be referred to as CN alignment in the following discussion, while the normal alignment is referred to as NC alignment (Figure 23). All in all, there were eight different setup for our AtPRF3-polyprolines simulations: N37L-CN, N37U-CN, N37L-NC, N37U-NC, N22L-CN, N22U-CN, N22L-NC and N22U-NC. The configurations are illustrated in Figure 23.

Fig 23. (a) Example of the initial binding positions of polyprolines to AtPRF3-N37. (b) Root means square deviation of distances between backbone atoms for homology model generated using SWISS-MODEL.

All molecular dynamics simulation were carried out using GROMACS version

5.1.2.^{150,198,240,241} All systems were equilibrated in canonical (NVT) ensemble and isothermal-isobaric (NPT) ensemble before production run (NPT) of 100 ns. CHARMM36 forcefield was used.²⁴² Particle Mesh Ewald method was employed for long range electrostatic calculation.¹⁵² The short range van der Waals and electrostatic cutoff were both set to 1.2 nm. LINCS algorithm was used to constrain bonds involving hydrogens in order to allow an integration timestep of 2 fs.¹⁵¹ The temperature of the system was set at 315 K which is slightly higher than the physiological temperature in order to accelerate sampling. For each of the eight binding model, three molecular dynamics simulations were initiated with different initial velocities for the atoms in order to sample a larger phase space. Secondary structures were calculated using “do_dssp” utilities based on the dictionary of secondary structures of protein (DSSP).¹⁶¹

5.2.2 MM/PBSA Binding Energy Analysis

In addition, we also carried out MM/PBSA calculations to estimate the binding affinity.²⁴³ In MM/PBSA calculations, the binding energy $E_{binding}$ is defined as:

$$E_{binding} = E_{vdW} + E_{elec} + E_{polar} + E_{apolar} - T\Delta S \quad (5.1)$$

where E_{vdW} and E_{elec} are the van der Waal’s interaction energy and electrostatic interaction energy respectively, both of which can be obtained from molecular dynamics simulations. The term E_{polar} refers to the polar solvation energy estimated through Poisson Boltzmann continuum solvent model.²⁴⁴ On the other hand, E_{apolar} is the apolar solvation energy estimated by using the Solvent Accessible Surface Area (SASA) model. Lastly, ΔS is the solute entropy. MM/PBSA has been applied successfully in a number of protein-protein interaction cases. We refer the interested reader to paper by Genheden et. al for detailed technical explanations of the MM/PBSA model.¹³⁷ A common approach in MM/PBSA is the calculation of the binding energy without considering the solute entropy. Our ligand consists of mainly prolines, which is known to be rigid and adopts PP-II helix as a secondary structure as discussed in section

5.1.²³¹ Thus in PBSA calculation, we can assume similar entropy contributions in all the configurations. This allows meaningful comparison between the different binding poses without explicitly calculating the entropy, which is in itself inaccurate and computationally challenging. The software “g_mmpbsa” with the default set of parameters was used for all binding energy calculations.¹⁵⁶ Bootstrap analysis with 2000 steps was used to calculate errors in free energy calculation.

5.3 Results

Firstly, in order to characterize the binding poses, we carried out the clustering of trajectories using gromos method.²⁴⁵ Structural clustering assigns conformations with similar root mean square deviations into clusters, allowing us to distinguish between structurally stable (high percentage population) conformations in simulations. Here, structures were fitted using backbone atoms of AtPRF3, and root mean square deviation (RMSD) was calculated using the full protein backbone. Cutoff was set at 0.25 nm. For each binding pose, three simulations were combined into a single trajectory before clustering. Frames every 50 ps were extracted to minimize statistical correlation. For AtPRF3-N37, the largest cluster for N37U-CN, N37L-NC and N37U-NC accounted for more than 20% of structures extracted. For N37U-CN, the first and second largest cluster accounted for more than 15% of structures extracted. In contrast, for AtPRF3-N22, the cluster members of N22L-CN and N22U-CN were more evenly distributed among the top three clusters in comparison. As such, we chose to analyse clusters that accounted for more than 15% of the total population (two clusters for N37L-CN, one cluster for N37U-CN, one cluster for N37L-NC, one cluster for N37U-NC and three clusters for both N22L-CN and N22U-CN).

For each cluster, we calculated the average distance between residues using heavy atoms (excluding hydrogen). The average distances between the residues pair of polyprolines (index 1-13) and AtPRF3-N37 (index 38-168) or AtPRF3-N22 (index 23-168) were then mapped onto contact heatmaps to visualize the binding interfaces. The representative structures in the clusters together with their respective contact heatmaps are illustrated in Figure 24 to Figure 26. We labelled the cluster index using numbers following the names of the configurations, i.e. N37L-CN-1 refers to the first (top) cluster, N37L-CN-2 refers to the second top cluster in N37L-CN simulations, and so on. Then, the binding free energies for the structural ensemble in each cluster were calculated using MM/PBSA method as described in section 5.2.2.

Next, we calculated the root means square deviation of distances (RMSD) between backbone atoms of the binding complexes in Figure S11 for each of the simulations. This enabled us to quantify the stability of binding complexes. It can be observed that the binding complexes were able to quickly achieve small fluctuation within 0.15 nm after a small initial period of equilibration. In general, all N37 configurations were able to achieve stable binding positions in our simulations. This was supported by the highly negative binding free energy (favorable binding) estimated through MM/PBSA tabulated in table S4 and table S5 (Supplementary information). On the other hand, while both N22L-CN and N22U-CN gave rise to stable binding complexes, we observed large changes in the binding position and ligand dissociation in N22L-NC and N22U-NC configuration (Supplementary Figure S12). Table S6 reveals that N22L-CN and N22U-CN resulted in stronger binding than N37 configuration, whereas N22L-NC and N22U-NC configuration resulted in positive binding free energy estimation (Table S7, supplementary information) which imply highly unstable complex. As a result, we will not be carrying out detailed analysis of N22L-NC and N22U-NC.

Fig 24. Figures on top show the representative structures for the largest cluster for simulation of AtPRF3-N37 in complex with polyprolines. Figures at the bottom show the corresponding contact maps between residues of AtPRF3-N37 (y-axis) and polyprolines (x-axis).

Fig 25. Figures on top show the representative structures for the largest cluster for simulation of AtPRF3-N22 in complex with polyprolines. Figures at the bottom shows the corresponding contact maps between residues of AtPRF3-N22 (y-axis) and polyprolines (x-axis).

Fig 26. Figures on top show the representative structures for the second and third largest cluster for simulation of AtPRF3-N22 in complex with polyprolines. Figures at the bottom show the corresponding contact maps between residues of AtPRF3-N22 (y-axis) and polyprolines (x-axis).

5.4 Discussions

5.4.1 Stability of AtPRF3-Polyprolines Complexes and Extended N-terminal α -helix

To discuss the validity of the homology modelling of AtPRF3, we look into two aspects. Firstly, the stabilities of the complexes were monitored by calculating the root-mean-square (RMS) deviation of the distances of backbone atoms. RMS-Distances plot for all the simulations are shown in Figure S11. We can see that in all the simulations, the complex stabilised after an initial period of increase in RMSD which is expected as the structures relax after the initial setup. Thus, the use of AtPRF1's crystal structure as a homology model can be justified. Secondly, it should be noted that PSIPRED actually predicted a section of random coil in between the N-terminal α -helix in AtPRF3-N22 and residue MET38 (First residue in AtPRF3-N37). The entire extended N-terminal sequences as an α -helix was modelled based on the rationale that it would result in a more compact structure to interact with polyprolines peptide; A coil would induce too much flexibility into the N-termini during the binding process, thereby it is unlikely to change the binding affinity. The binding of a ligand may also stabilize the secondary structures.

Nevertheless, in order to test the structural stability of our modelled α -helix, we calculated the secondary structures of the extended sequences using the final 25 ns simulation from each of the three repeats for N22L-CN and N22U-CN using the dictionary of secondary structures described in section 5.2.1. The percentage of α -helix adopted by the N-terminal extension is tabulated in table 7. All simulations were able to maintain high amount of α -helix content (more than 50%), thus supporting our modelling of the extension. Visual inspection of structures in Figure 25 and 26 supports the findings.

Table 7. Table shows the α -helix content of the constructed N-terminal extension **QRRSRAKVKKKKKTN**. Percentage content was calculated as $\frac{\text{Number of } \alpha\text{-helix}}{15}$ where 15 is the length of the extension. Errors were estimated from the standard deviation during the last 25 ns of each run.

		Percentage α -helix	Standard Deviation
N22L-CN	Run 1	71.847 %	12.780 %
	Run 2	61.942 %	6.480 %
	Run 3	63.875 %	4.436 %
N22U-CN	Run 1	66.002 %	3.194 %
	Run 2	52.509 %	3.965 %
	Run 3	66.643 %	0.442 %

5.4.2 CH- π and Electrostatic Interactions

The recognition of polyprolines is one of the most important biological functions. Notably, the Src Homology 3 (SH3) domain, one of the most well studied proteins, contains a binding groove that allows peptides with sequence motif PXXP to attach on (where X is any amino acid).²⁴⁶ WW domain of brain protein F65 is also known to recognize polyprolines.²⁴⁷ Besides the importance of polyprolines' rigidity in forming PP-II helix allowing it to "slot in" to the binding sites, a common feature found among proteins that recognise polyprolines is the presence of aromatic residues such as tyrosine.^{231,246-248} The interaction between aromatic residues and prolines ring is known as the CH- π interaction, named after the fact that this interaction happens between the π aromatic face and the polarized C-H bonds of proline residues.²⁴⁹ CH- π interactions have been investigated in the study of mammalian profilin 2a in complex with two polyprolines ligand by Kursula et. al.²²⁵ Besides, hydrogen bond donations by tryptophan, tyrosine and asparagine were discovered to be critical in ligand recognition in the same study.

In this context, a number of aromatic and polar residues (hydrogen bonds donor) can be found in SER39, TRP40, TYR43, HIS47, TRP70, ASN158, TYR162 and SER166 on AtPRF3 studied in our simulations. SER39, TRP40, TYR43 and HIS47 are located at the N-terminal helix, while TRP70 is located at one of the loops close to the U binding mode. The remaining TYR162 and SER166 are located at the C-terminal helix. From contact heatmap in Figure 24 and 25, we can see that these

residues are indeed often involved in the binding interface in our simulation, consistent with the literature. In particular, aromatic residues TRP40, TYR43, HIS47 and TYR162 were involved universally in all binding modes with the exception of N37L-CN-1, confirming the importance of CH- π interactions. Other than that, while not universally present in all binding modes, polar residues SER39, ASN158 and SER166 were also found to be involved in many configurations (highlighted in blue frames from Figure 24 to Figure 26).

A notably unique binding mode was found in N37L-CN-1 where we observed relatively few residues involved in the binding, yet the complex was seen to be highly stable with strong binding affinity. A contact heatmap (figure 24) shows that the positively charged residue ARG1 on polyprolines binds very strongly to the negatively charged residues ASP46 and ASP51 on AtPRF3 in both N37L-CN binding modes. Similar electrostatic interaction was also observed in N37U-NC binding mode with ASP66. Interestingly, in the N22 binding mode, besides observing electrostatic attractions involving ASP46 and ASP51 in N22L-CN-2, N22L-CN-3 and N22U-CN-2, we found interactions between polyprolines-ARG1 and AtPRF3-GLU165. In addition, Figure 25 shows that the N-terminal extension has a number of contact hotspots with the second half of polyprolines sequence. These additional contact hotspots are the main reason of stronger binding in AtPRF3-N22 as we will discuss later.

In short, we found electrostatic interactions to be another critical factor in recognizing polyprolines region containing positively-charged residues. This type of recognition mechanism was also found in SH3-PXXP binding.^{5,246} Indeed, by examining the full sequences of FH1 domain of *arabidopsis thaliana* (Table S8, supplementary information), we identified a few potential polyprolines sequences: **DSPPPSPPSPPPLPK** LP, YPSSPPPPSP, **RRVYPPPPATAPPTR**, **RPPSLTPP** and lastly the sequence studied in experiment **RVPPPPPPPPPLP**.²⁵⁰ Except for the second sequence, it can be observed that the polyprolines sequences are flanked by charged residue aspartate or arginine. The orientation of polyprolines can thus be dependant upon the position

of charged residues.

5.4.3 MM/PBSA Binding Free Energy

To quantify the binding affinity, we used MM/PBSA method to estimate binding energy between profilin and polyprolines in our simulation. As we can see, the binding free energies are negative, indicating favourable binding. In general, from table S4 to table S7, CN binding modes were observed to have stronger binding affinity compared to NC binding modes. The binding free energies for AtPRF3-N22 are higher than those of AtPRF3-N37. We note that MM/PBSA has inherent limitations in predicting the absolute binding free energy and should only be relied to predict the relative strength of binding free energy between similar proteins. In this context, our results are consistent with the results from biochemical experiments, where N22 has a dissociation constant, K_d value of $70 \pm 6.31 \mu M$, almost twice as much stronger than that of N37 which has a K_d value of $175.6 \pm 3.78 \mu M$ (Manuscript in preparation¹⁴³). While the absolute binding free energies are less meaningful, the individual contributions from different energy terms will help us understand differences between the binding modes.

By decomposing $E_{binding}$ into its individual component, we can examine the causes resulting in stronger binding. N37L-CN was seen to have low Van der Waals force due to its loose binding. This was however highly compensated by strong electrostatic energy due to its interaction with ASP46 and ASP51 discussed above. Compared to it, N37U-CN packed polyprolines in a tighter fashion resulting in stronger Van der Waal's energy. It has less electrostatic contribution but at the same time less polar solvation penalty due to tighter packing of the hydrophobic polyprolines. In fact, the combined contribution of E_{elec} and E_{polar} is critical to the complexes. The values are tabulated in table S4 to table S7. As such, we can observe that in general, NC binding mode has a comparatively low combined contribution of E_{elec} and E_{polar} . Contact heatmaps showed that the major differences between CN and NC binding modes are due to lack of interaction hotspots around ASP46 and ASP51. Besides, as the polyprolines are

uncapped, its C-termini is negatively-charged. This may cause a large electrostatic penalty as the C-termini is close to ASP46 and ASP51 in the NC configuration.

Next, we examined the binding free energy estimated for N22L-CN and N22U-CN. By looking at the individual contribution again, it can be noticed that N22-polyprolines complexes have a much higher electrostatic energy component than N37-polyprolines due to the N-terminal extension which contains many positively-charged lysine residues. The combined contribution of E_{elec} and E_{polar} is higher as well when comparing N37L-CN against N22L-CN and N37U-CN against N22U-CN. This is due to the fact the the N-terminal extension is highly hydrophilic and hence its polar solvation penalty is not large enough to negate the highly favourable electrostatic interaction between itself and polyprolines. On the other hand, N22L-NC and N22U-NC both orient ARG1 of polyprolines close to the positively-charged N-terminal extension, causing strong electrostatic repulsion. Consequently, stable binding modes were not possible in N22L-NC and N22U-NC.

Fig S10. Prediction of secondary structures for AtPRF3-N22 using PSIPRED web-server.

Fig S11. RMS Deviation of distance of the backbone atoms of the complexes during simulation for (a) N37L-CN (b) N37U-CN (c) N37L-NC (d) N37U-NC (e) N22L-CN (f) N22U-CN

Fig S12. (a) Peptide dissociation in N22L-NC. Other than dissociating from the binding groove, N-termini can also be seen to have rotated to the opposite orientation compared to the initial position (figure 23). (b) Unfolded N-terminal helix in N22U-NC. Cyan color represents the N-terminal extension. Blue sphere indicates the N-termini of polyprolines. Figures are generated by author using PyMOL.¹

Table S4. Binding free energy estimated using MM/PBSA for N37L-CN and N37U-CN configurations. Unit for energy is kJ/mol.

		Percentage Population	E_{vdW}	E_{elec}	E_{polar}	E_{SASA}	$E_{binding}$	$E_{elec} + E_{polar}$
N37L-CN	Cluster 1	10.78%	-60.119±0.646	-730.081±1.807	485.480±1.795	-10.480±0.072	-315.082±1.565	-244.601
	Cluster 2	8.50%	-131.145±1.236	-577.436±3.795	463.421±4.355	-17.931±0.111	-263.110±2.044	-114.015
N37U-CN	Cluster 1	23.01%	-194.794±0.912	-514.057±4.334	384.410±5.725	-23.771±0.105	-348.273±2.983	-129.647

Table S5. Binding free energy estimated using MM/PBSA for N37L-NC and N37U-NC configurations. Unit for energy is kJ/mol.

		Percentage Population	E_{vdW}	E_{elec}	E_{polar}	E_{SASA}	$E_{binding}$	$E_{elec} + E_{polar}$
N37L-NC	Cluster 1	21.73%	-158.157±1.155	-94.332±2.928	69.200±2.399	-19.202±0.139	-202.570±2.706	-25.132
N37U-NC	Cluster 1	24.87%	-244.412±2.060	-320.555±3.798	348.167±4.462	-29.657±0.243	-246.452±2.866	27.612

Table S6. Binding free energy estimated using MM/PBSA for N22L-CN and N22U-CN configurations. Unit for energy is kJ/mol.

		Percentage Population	E_{vdW}	E_{elec}	E_{polar}	E_{SASA}	$E_{Binding}$	$E_{elec} + E_{polar}$
N22L-CN	Cluster 1	7.03%	-240.736±1.894	-1181.255±13.813	832.933±12.785	-32.395±0.241	-621.575±3.097	-348.322
	Cluster 2	6.43%	-194.535±1.747	-1047.883±7.199	798.896±9.641	-27.558±0.172	-470.974±4.362	-248.987
	Cluster 3	5.65%	-180.469±2.356	-1164.068±14.068	915.254±14.931	-27.195±0.182	-456.706±4.264	-248.814
N22U-CN	Cluster 1	6.83%	-280.701±3.534	-857.633±17.893	649.905±15.613	-34.745±0.338	-523.462±7.686	-207.728
	Cluster 2	5.37%	-252.647±3.577	-988.071±15.482	744.312±12.586	-33.369±0.333	-529.498±5.866	-243.759
	Cluster 3	5.05%	-259.703±2.412	-1280.212±10.868	916.681±9.894	-36.586±0.247	-659.679±3.341	-363.531

Table S7. Binding free energy estimated using MM/PBSA for N22L-NC and N22U-NC configurations. Unit for energy is kJ/mol.

		Percentage Population	E_{vdW}	E_{elec}	E_{polar}	E_{SASA}	$E_{Binding}$	$E_{elec} + E_{polar}$
N22L-NC	Cluster 1	10.43%	-257.220±1.453	156.395±5.766	412.979±4.624	-31.993±0.173	280.346±3.593	569.374
	Cluster 2	7.10%	-176.422±2.485	-327.295±6.159	558.120±8.610	-26.026±0.277	28.101±5.127	230.825
	Cluster 3	3.85%	-211.903±1.969	148.179±11.246	373.345±10.719	-25.768±0.208	283.835±7.975	521.524
N22U-NC	Cluster 1	11.71%	-281.150±1.487	107.384±6.318	328.621±3.812	-32.725±0.134	122.277±3.725	436.005
	Cluster 2	7.52%	-243.020±1.741	250.213±9.749	168.278±5.333	-29.381±0.194	146.083±6.012	418.491
	Cluster 3	3.83%	-189.393±1.966	55.689±14.769	252.459±10.318	-23.443±0.214	95.428±7.834	308.148

Table S8. Full FASTA sequence of formin FH1 domain of *arabidopsis thaliana*

MLFFLFFFYLLSSSSDLVFADRRVLHEPFFFPIDSPPPSPSPPLPKLPFSSTTPPSSSDPNASFFFP
YPSSPPPPSPASFAFPANISLIVPHATKSPNSKLLIVAISAVSSAALVALLIALLYWRRSKRNQDL
NFSDDSKTYTTDSSRRVYPPPPATAPPTRRNAEARSKQRTTTSSTNNNSSEFLYLGTMVNRGIDEQSLS
NNGSSSRKLESPDLQPLPPLMKRSFRLNPDVGSIGEEDEEDEFYSPRGSQSGREPLNRVGLPGQNPRSVN
NDTISCSSSSSGSPGRSTFISISPSMSPKRSEPKPPVISTPEPAELTDYRFVRSPSLSLASLSSGLKNSD
EVGLNQIFRSPTVTSLTTSPENKKKENSPLSSTSTSPERRPNDTPEAYLRSPSHSSASTSPYRCFQKSPE
VLPAFMSNLRQGLQSLLSSPSNSHGGQGFLKQLDALRSRSPSSSSSVCSSEKASHKSPVTS PKLSSR
NSQSLSSSPDRDFSHSLDVSPRISNISPQILQSRVPPPPPPPPPLPLWGRRSQVTTKADTISRPPSLTTP
SHPFVIPSENLPVTSSPMETPETVCASEAAEETPKPKLKALHWDKVRASSDREMVDHLRSSSFKLDEEM
IETLFVAKSLNNKPNQSQTTTPRCVLPSPNQENRVLDPKKAQNIAILLRALNVTIEEVCEALLEGNADTLG
TELLESLLKMAPTKEEERKLKAYNDDSPVKLGHA EKFLKAMLDIPFAFKRVDAMLYVANFESEVEYLKKS
FETLEAAACEELRNSRMFLKLLEAVLKTGNRMNVGTNRGDAHAFKLDTLKLV DVKGADGKTTLHFV VQE
IIRAEGTRLSGNNTQTDDIKCRKLGLQVVSSLCSELSNVKKAAMDSEVLSSYVSKLSQGI AKINEAIQV
QSTITEESNSQRFSESMKTFLKRAEEEIIRVQAQESVALSLVKEITEYFHGNSAKEEAHPFRIFLVVRDF
LGVVDRVCKEVGMINERTMVSSAHKFVPVNPMPQPLPGLVGRRQSSSSSSSSTSSSDEDEHNSISLV
S

6 Conclusion and Future Outlook

In this thesis, three types of protein-protein interactions were explored: cross-species proteins aggregation, self-assembly of short peptides into hydrogel, and polyprolines recognition mechanism. The first two topics can also be classified together as aggregation phenomenon, albeit in different contexts. In cross-species aggregation, we sought to understand how amylin and PrP106-126 can aggregate despite being structurally and sequentially distinct. The two peptides were studied by investigating into their dimerization process. In the original experiment by Gal et al. which inspired our study,¹³⁹ the possibility of heterogeneous fibril could not be validated. In our study, we found that the two peptides were able to form highly polymorphic complex. Despite this polymorphic nature, energy minima with highly ordered β -sheets could be found. This implied that heterogeneous fibril is highly probable.

In addition to that, our molecular dynamics simulation provided a way to investigate interactions between amino acids involved in aggregation between two distinct amyloidogenic proteins. We were able to pinpoint interaction hotspots using our data. Interestingly, some of the residues traditionally found to be important in the same-species aggregations of amylin and PrP106-126 were also highly involved in the cross-species aggregation. The pathological significance of cross-species aggregation is a relatively new aspect of protein-protein interactions. Most of the focus currently is on A β and amylin due to the abundance of existing literatures and direct identification of type II diabetes mellitus as a risk factor of Alzheimer's disease. While amylin and PrP106-126 are physiologically unrelated, it is important to understand the cross-talk between these two proteins as the mechanism is potentially applicable to other amyloidogenic proteins.

The aggregation of peptides can also result in highly ordered polymeric networks and produce hydrogels. In Chapter 4 of this thesis, we used molecular dynamics to understand the self-assembly of peptide GV8 and GL8. These two peptides were part

of the sucker ring teeth (SRT) proteins found in squids. They were identified by our experimental collaborators to form structures with a high percentage of β -sheet. Peptide GV8 in particular forms hydrogel whereas peptide GL8 does not, despite only having a single amino acid difference in their sequences. This amino acid difference gave rise to different hydrophobicity (GL8 > GV8). We used an approach inspired by Frederix et. al in which coarse-grained modelling using Martini force field was employed to look at the supramolecular assembly of peptides.¹³⁰ It was found that peptide GL8 adopted a more elongated fibril-like structure whereas GV8 adopted a more disc-like structures. Simulations in high concentration yielded supramolecular assembly consist of the peptides in fibrils-like arrangement possibly resembling that of a hydrogel.

Furthermore, we extended our study by using metadynamics, an enhanced sampling method to study the self-assembly process in fine-grained molecular dynamics. This allowed us to look into the interactions in a more accurate (atomistic) manner, despite our limitation to a tetramer configuration due to computational capability. Both GV8 and GL8 sampled similar secondary structures in the ensemble average quantities. However, significant differences were identified for structures with low free energy. In these structures, peptide GL8 was found to possess more β -sheets secondary structure than peptide GV8. Visualization of clusters obtained in the simulations displayed similar characteristic with peptide GL8 forming a more highly ordered structures. In a nutshell, our simulations suggest a larger proportion of β -sheets at in small oligomers (tetramer in our case) is potentially unfavorable for the formation of hydrogel.

The collaboration in this project is still ongoing. Experiments for a few other peptides with single amino acid differences (As in GV8 and GL8, where substitutions are made with alanine (GA8), phenylalanine (GF8), serine (GS8) and lysine (GK8)) are being planned. We hope to have a complete picture of the effect of hydrophobicity and electrostatic interactions in hydrogel formation of peptides from sucker ring teeth proteins. Eventually, this project will provide insights into structural determinants

of hydrogel and potentially inspire the design of novel hydrogel using peptides with similar sequences.

Next, Chapter 5 described our work in the polyprolines recognition by profilin, a mechanism vital for regulation of actin polymerization. Profilin has been known to bind to both actin and formin using two different binding sites. Binding of formin involves the various polyprolines sites in the formin's FH1 domain. In collaboration with our experimental collaborators, we modelled the binding complex of profilin (AtPRF3) and a polyprolines segment from formin FH1 (AtFH1) of plant species *Arabidopsis thaliana*. Our main contribution included complementing the biochemical assays with a structural model of AtPRF3-N37 in complex with polyprolines as the complex could not be solved through X-ray crystallography. Literature suggests two potential binding orientations in which we were able to test using molecular dynamics simulations and found one to be more stable than the other.¹⁴³

In addition, we built models of a longer version of profilin (AtPRF3-N22) containing an extended N-termini with a lot of charged amino acids compared to AtPRF3-N37. In existing literatures, studies were done on short version of profilin similar to AtPRF3-N37. However, it is possible for profilins with longer N-terminus to be expressed by cells in nature. Thus, it is important to understand the interactions between AtPRF3-N22 and polyprolines. Using secondary structures prediction, the extension was modelled as an α -helix. Notably, this model was validated to be stable in our molecular dynamics simulations. Using MM/PBSA method, the binding free energy of AtPRF3N22-polyprolines was found to be stronger than that of AtPRF3N37-polyprolines, consistent with the results obtained by our experimental collaborators. We were able to uncover the role of electrostatic interactions in stabilizing the binding complex as well as determining the binding orientation of polyprolines.

Protein-protein interactions will continue to be an important topic in our quest for complete understanding of the physiological significance of proteins in living organisms. Some literature strive to provide a macroscopic understanding by means

such as looking at genomic data bank and carrying out network analysis of known protein-protein interactions.²⁵¹⁻²⁵⁴ It is important to note however that the microscopic understanding of interactions at the atomistic scale is equally crucial in many fields. Important examples include drugs and biomaterials design in which knowing the roles of amino acids in structural models is often the first step towards the goal. Molecular simulations serve to fill in this role and it has been the theme in this thesis. By covering a range of topics in this direction, this thesis aims provide a valuable foundation for more extensive and exhaustive studies in the future.

Ultimately, while significant, we acknowledge that the works in this thesis are just a mere glimpse into the vast amount of protein-protein interactions. At the same time, despite being increasingly common as a tool to complement experimental studies of proteins, computational model has its shortcomings. Our work in cross-species aggregation for example, used high local concentration to mimic the effect of lipid bilayers. This was mainly due to lack of established protocol for enhanced sampling of proteins together with lipids. A conventional simulations of such a complicated system would require prohibitively expensive computational time for reasonable convergence as it is a protein folding problem (Chapter 1). In the future, we hope that development of sophisticated protocols and continuous improvement in computational speed would enable more accurate and extensive simulation in simulating real physical interactions of peptides on lipid bilayer.

Designing versatile biomaterials using peptides has long been the goals of many material scientists. The continuous refinement of coarse-grained model such as Martini model used in this thesis will provide fast and accurate screening of large number of peptides that can potentially form interesting nanostructures. As we saw in our work, there were still limitations to Martini force field in modelling hydrophobicity such that we could not find differences as significant as what was seen in experiments. Nevertheless, what we found showed promising potential and would provide valuable informations for researchers developing the force field to improve on.

In our study of AtPRF3 in complex with polyprolines, the N-terminal extension is not unique to just AtPRF3. There are other profilins containing such extension, albeit with different types of residues. This N-terminal extension has not been explored in literature to the best of our knowledge. Our work aims to highlight the existence of this N-terminal extension as well as its roles in the recognition of polyprolines compared to profilins without the extension. In the future, more experiments can be done on other species of profilin with N-terminal extension and molecular dynamics simulations may continue to provide insights to the specific interactions involved similar what we found regarding the electrostatic interactions.

7 List of Publications

1. Chua, K. P., Chew, L. Y., & Mu, Y. (2016). Replica Exchange Molecular Dynamics Simulation of Cross-Fibrillation of IAPP and PrP106-126. *Proteins: Structure, Function, and Bioinformatics*, (April), 113. <https://doi.org/10.1002/prot.25060>
2. Sun, H., Qiao, Z., Chua, K. P., Tursic, A., Koh, S. H., Liu, X., Gao, Y. G., Mu, Y., Hou, X., Miao, Y. Negative Regulation of Formin-Mediated Actin Assembly by Profilin 3 is Positively Involved in Arabidopsis Innate Immunity. (Manuscript submitted, in revision)

8 References

- [1] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.
- [2] Wikimedia Commons. File:beta sheet bonding antiparallel-color.svg — wikimedia commons, the free media repository, 2010. [Online; accessed 31-July-2017].
- [3] Wikimedia Commons. File:beta sheet bonding parallel-color.svg — wikimedia commons, the free media repository, 2010. [Online; accessed 31-July-2017].
- [4] Thorsten Lührs, Christiane Ritter, Marc Adrian, Dominique Riek-Loher, Bernd Bohrmann, Heinz Döbeli, David Schubert, and Roland Riek. 3D structure of Alzheimer’s amyloid-beta(1-42) fibrils. *Proceedings of the National Academy of Sciences of the United States of America*, 102(48):17342–7, nov 2005.
- [5] S Feng, J. Chen, H Yu, J. Simon, and S. Schreiber. Two binding orientations for peptides to the Src SH3 domain: development of a general model for SH3-ligand interactions. *Science*, 266(5188):1241–1247, nov 1994.
- [6] E A Brucker. Genetically crosslinked hemoglobin: a structural study. *Acta crystallographica. Section D, Biological crystallography*, 56(Pt 7):812–6, jul 2000.
- [7] Ravi Prakash Reddy Nanga, Jeffrey R. Brender, Subramanian Vivekanandan, and Ayyalusamy Ramamoorthy. Structure and membrane orientation of IAPP in its natively amidated form at physiological pH in a membrane environment. *Biochimica et Biophysica Acta - Biomembranes*, 1808(10):2337–2342, 2011.
- [8] Jakob T. Nielsen, Morten Bjerring, Martin D. Jeppesen, Ronnie O. Pedersen, Jan M. Pedersen, Kim L. Hein, Thomas Vosegaard, Troels Skrydstrup, Daniel E. Otzen, and Niels C. Nielsen. Unique Identification of Supramolecular Structures in Amyloid Fibrils by Solid-State NMR Spectroscopy. *Angewandte Chemie*, 121(12):2152–2155, mar 2009.
- [9] R Zahn, a Liu, T Lührs, R Riek, C von Schroetter, F López Garcia, M Billeter, L Calzolari, G Wider, and K Wüthrich. NMR solution structure of the human prion protein. *Proceedings of the National Academy of Sciences*, 97:145–150, 2000.
- [10] Karen J. Knaus, Manuel Morillas, Wieslaw Swietnicki, Michael Malone, Witold K. Surewicz, and Vivien C. Yee. Crystal structure of the human prion

- protein reveals a mechanism for oligomerization. *Nature Structural Biology*, 8(9):770–774, sep 2001.
- [11] Yong-Xing He, Nan-Nan Zhang, Wei-Fang Li, Ning Jia, Bao-Yu Chen, Kang Zhou, Jiahai Zhang, Yuxing Chen, and Cong-Zhao Zhou. N-Terminal Domain of Bombyx mori Fibroin Mediates the Assembly of Silk in Response to pH Decrease. *Journal of Molecular Biology*, 418(3-4):197–207, may 2012.
- [12] Noga Gal, Ahiud Morag, Sofiya Kolusheva, Roland Winter, Meytal Landau, and Raz Jelinek. Lipid Bilayers Significantly Modulate Cross-Fibrillation of Two Distinct Amyloidogenic Peptides. *Journal of the American Chemical Society*, 135(36):13582–13589, sep 2013.
- [13] Patrice Koehl. Protein Structure Classification. pages 1–55. John Wiley & Sons, Inc., feb 2006.
- [14] K.U. Linderstrøm-Lang. *Lane Medical Lectures: Proteins and Enzymes*. Med. sciences: Stanford University publications / Univ. series. Stanford University Press, 1952.
- [15] Peter E Wright and H.Jane Dyson. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, 293(2):321–331, oct 1999.
- [16] A K Dunker, J D Lawson, C J Brown, R M Williams, P Romero, J S Oh, C J Oldfield, A M Campen, C M Ratliff, K W Hipps, J Ausio, M S Nissen, R Reeves, C Kang, C R Kissinger, R W Bailey, M D Griswold, W Chiu, E C Garner, and Z Obradovic. Intrinsically disordered protein. *Journal of molecular graphics & modelling*, 19(1):26–59, 2001.
- [17] Oliver C. Redfern, Benoit Dessailly, and Christine A. Orengo. Exploring the structure and function paradigm. *Current Opinion in Structural Biology*, 18(3):394–402, jun 2008.
- [18] Peter E. Wright and H. Jane Dyson. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology*, 16(1):18–29, dec 2014.
- [19] Robert Zwanzig, Attila Szabo, and Biman Bagchi. Levinthal’s paradox. *Proceedings of the National Academy of Sciences*, 89(1):20–22, 1992.

-
- [20] P E Leopold, M Montal, and J N Onuchic. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proceedings of the National Academy of Sciences*, 89(18):8721–8725, sep 1992.
- [21] Joseph D. Bryngelson, José Nelson Onuchic, Nicholas D. Socci, and Peter G. Wolynes. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins: Structure, Function, and Genetics*, 21(3):167–195, mar 1995.
- [22] José Nelson Onuchic, Zaida Luthey-Schulten, and Peter G. Wolynes. THEORY OF PROTEIN FOLDING: The Energy Landscape Perspective. *Annual Review of Physical Chemistry*, 48(1):545–600, oct 1997.
- [23] Claudio Soto. Unfolding the role of protein misfolding in neurodegenerative diseases. *Nature Reviews Neuroscience*, 4(1):49–60, jan 2003.
- [24] Fabrizio Chiti and Christopher M. Dobson. Protein Misfolding, Functional Amyloid, and Human Disease. *Annual Review of Biochemistry*, 75(1):333–366, jun 2006.
- [25] M. Paul Murphy and Harry LeVine. Alzheimer’s Disease and the Amyloid- β Peptide. *Journal of Alzheimer’s Disease*, 19(1):311–323, jan 2010.
- [26] I. W. Hamley. The Amyloid Beta Peptide: A Chemist’s Perspective. Role in Alzheimer’s and Fibrillization. *Chemical Reviews*, 112(10):5147–5192, oct 2012.
- [27] J. Hardy. The Amyloid Hypothesis of Alzheimer’s Disease: Progress and Problems on the Road to Therapeutics. *Science*, 297(5580):353–356, jul 2002.
- [28] Franklin H. Epstein, Jo W.M. Höppener, Bo Ahrén, and Cornelis J.M. Lips. Islet Amyloid and Type 2 Diabetes Mellitus. *New England Journal of Medicine*, 343(6):411–419, aug 2000.
- [29] J. Götz, L. M. Ittner, and Y.-A. Lim. Common features between diabetes mellitus and Alzheimer’s disease. *Cellular and Molecular Life Sciences*, 66(8):1321–1325, apr 2009.
- [30] Aphrodite Kapurniotu. Amyloidogenicity and cytotoxicity of islet amyloid polypeptide. *Biopolymers*, 60(6):438–459, 2001.
- [31] Peter T. Lansbury. In pursuit of the molecular structure of amyloid plaque: new technology provides unexpected and critical information. *Biochemistry*, 31(30):6865–6870, aug 1992.

- [32] N. A. Peppas, Y. Huang, M. Torres-Lugo, J. H. Ward, and J. Zhang. Physico-chemical Foundations and Structural Design of Hydrogels in Medicine and Biology. *Annual Review of Biomedical Engineering*, 2(1):9–29, aug 2000.
- [33] Tetsuo Asakura, Juming Yao, Tsutomu Yamane, Kosuke Umemura, and Anne S. Ulrich. Heterogeneous Structure of Silk Fibers from *Bombyx mori* Resolved by ^{13}C Solid-State NMR Spectroscopy. *Journal of the American Chemical Society*, 124(30):8794–8795, jul 2002.
- [34] Adam Blazas, Steven P. Armes, and Anthony J. Ryan. Self-Assembled Block Copolymer Aggregates: From Micelles to Vesicles and their Biological Applications. *Macromolecular Rapid Communications*, 30(4-5):267–277, feb 2009.
- [35] Anupama Lakshmanan, Shuguang Zhang, and Charlotte A.E. Hauser. Short self-assembling peptides as building blocks for modern nanodevices. *Trends in Biotechnology*, 30(3):155–165, mar 2012.
- [36] Zhiqiang Yu, Quan Xu, Chenbo Dong, Su Lee, Liqian Gao, Yiwen Li, Mathew D’Ortenzio, and Jun Wu. Self-Assembling Peptide Nanofibrous Hydrogel as a Versatile Drug Delivery Platform. *Current Pharmaceutical Design*, 21(29):4342–4354, oct 2015.
- [37] W. John Albery and Jeremy R. Knowles. Evolution of enzyme function and the development of catalytic efficiency. *Biochemistry*, 15(25):5631–5640, dec 1976.
- [38] Peer Bork, Chris Sander, and Alfonso Valencia. Convergent evolution of similar enzymatic function on different protein folds: The hexokinase, ribokinase, and galactokinase families of sugar kinases. *Protein Science*, 2(1):31–40, dec 2008.
- [39] S H Bhosale, M B Rao, and V V Deshpande. Molecular and industrial aspects of glucose isomerase. *Microbiological reviews*, 60(2):280–300, jun 1996.
- [40] Ana C. Ramos-Valdivia, Robert van der Heijden, and Robert Verpoorte. Isopentenyl diphosphate isomerase: a core enzyme in isoprenoid biosynthesis. A review of its biochemistry and function. *Natural Product Reports*, 14(6):591, 1997.
- [41] David F. Waugh. Protein-Protein Interactions. pages 325–437. 1954.
- [42] Susan Jones and Janet M Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93(1):13–20, jan 1996.

-
- [43] G K Ackers and F R Smith. The Hemoglobin Tetramer:A Three-State Molecular Switch for Control of Ligand Affinity. *Annual Review of Biophysics and Biophysical Chemistry*, 16(1):583–609, jun 1987.
- [44] J Schlessinger. SH2/SH3 signaling proteins. *Current opinion in genetics & development*, 4(1):25–30, feb 1994.
- [45] Walter Witke. The role of profilin complexes in cell motility and other cellular processes. *Trends in Cell Biology*, 14(8):461–469, aug 2004.
- [46] Irwin D Kuntz. Structure-based strategies for drug design and discovery. *Science*, pages 1078–1082, 1992.
- [47] Pierre De Meyts and Jonathan Whittaker. Structural biology of insulin and IGF1 receptors: implications for drug design. *Nature Reviews Drug Discovery*, 1(10):769–783, oct 2002.
- [48] Christopher W Murray and Tom L Blundell. Structural biology in fragment-based drug design. *Current Opinion in Structural Biology*, 20(4):497–507, aug 2010.
- [49] Michael Levitt. Molecular dynamics of native protein. *Journal of Molecular Biology*, 168(3):595–617, aug 1983.
- [50] A Kolinski and J Skolnick. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins*, 18(4):338–52, apr 1994.
- [51] U H Hansmann and Y Okamoto. New Monte Carlo algorithms for protein folding. *Current opinion in structural biology*, 9(2):177–83, apr 1999.
- [52] Rajdeep K. Grewal Roy and Soumen. Modeling proteins as residue interaction networks. *Protein & Peptide Letters*, 22:923–933, 2015.
- [53] J Andrew McCammon, Bruce R Gelin, and Martin Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585–590, 1977.
- [54] Krista Conger. Michael levitt wins nobel prize in chemistry, 10 2013.
- [55] V. S. Pande and D. S. Rokhsar. Molecular dynamics simulations of unfolding and refolding of a β -hairpin fragment of protein G. *Proceedings of the National Academy of Sciences*, 96(16):9062–9067, aug 1999.

-
- [56] A. Liwo, M. Khalili, and H. A. Scheraga. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proceedings of the National Academy of Sciences*, 102(7):2362–2367, feb 2005.
- [57] Matthew J. Harvey and Gianni De Fabritiis. High-throughput molecular dynamics: the powerful new tool for drug discovery. *Drug Discovery Today*, 17(19-20):1059–1062, oct 2012.
- [58] I. Buch, T. Giorgino, and G. De Fabritiis. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proceedings of the National Academy of Sciences*, 108(25):10184–10189, jun 2011.
- [59] Ehud Gazit. Mechanisms of amyloid fibril self-assembly and inhibition. *FEBS Journal*, 272(23):5971–5978, nov 2005.
- [60] Tao Yu and George C. Schatz. Free-Energy Landscape for Peptide Amphiphile Self-Assembly: Stepwise versus Continuous Assembly Mechanisms. *The Journal of Physical Chemistry B*, 117(45):14059–14064, nov 2013.
- [61] Bogdan Barz and Brigita Urbanc. Minimal Model of Self-Assembly: Emergence of Diversity and Complexity. *The Journal of Physical Chemistry B*, 118(14):3761–3770, apr 2014.
- [62] Jakob P. Ulmschneider, Martin B. Ulmschneider, and Alfredo Di Nola. Monte Carlo vs Molecular Dynamics for All-Atom Polypeptide Folding Simulations. *The Journal of Physical Chemistry B*, 110(33):16733–16742, aug 2006.
- [63] William L. Jorgensen and Julian Tirado-Rives. Monte Carlo vs Molecular Dynamics for Conformational Sampling. *The Journal of Physical Chemistry*, 100(34):14508–14513, jan 1996.
- [64] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1-2):141–151, nov 1999.
- [65] Yuguang Mu, Ye Yang, and Weixin Xu. Hybrid Hamiltonian replica exchange molecular dynamics simulation method employing the Poisson-Boltzmann model. *The Journal of chemical physics*, 127(8):084119, aug 2007.
- [66] Weixin Xu, Tingfeng Lai, Ye Yang, and Yuguang Mu. Reversible folding simulation by hybrid Hamiltonian replica exchange. *Journal of Chemical Physics*, 128(17):175105, may 2008.

- [67] Alessandro Laio and Francesco L Gervasio. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics*, 71(12):126601, 2008.
- [68] David E. Shaw, Jack C. Chao, Michael P. Eastwood, Joseph Gagliardo, J. P. Grossman, C. Richard Ho, Douglas J. Lerardi, István Kolossváry, John L. Klepeis, Timothy Layman, Christine McLeavey, Martin M. Deneroff, Mark A. Moraes, Rolf Mueller, Edward C. Priest, Yibing Shan, Jochen Spengler, Michael Theobald, Brian Towles, Stanley C. Wang, Ron O. Dror, Jeffrey S. Kuskin, Richard H. Larson, John K. Salmon, Cliff Young, Brannon Batson, and Kevin J. Bowers. Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM*, 51(7):91, jul 2008.
- [69] David E. Shaw, Kevin J. Bowers, Edmond Chow, Michael P. Eastwood, Douglas J. Ierardi, John L. Klepeis, Jeffrey S. Kuskin, Richard H. Larson, Kresten Lindorff-Larsen, Paul Maragakis, Mark A. Moraes, Ron O. Dror, Stefano Piana, Yibing Shan, Brian Towles, John K. Salmon, J. P. Grossman, Kenneth M. Mackenzie, Joseph A. Bank, Cliff Young, Martin M. Deneroff, and Brannon Batson. Millisecond-scale molecular dynamics simulations on Anton. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis - SC '09*, number c, page 1, New York, New York, USA, 2009. ACM Press.
- [70] David E Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science*, 330(6002):341–346, oct 2010.
- [71] S. Piana, K. Lindorff-Larsen, and D. E. Shaw. Protein folding kinetics and thermodynamics from atomistic simulation. *Proceedings of the National Academy of Sciences*, 109(44):17845–17850, oct 2012.
- [72] S. Piana, K. Lindorff-Larsen, and D. E. Shaw. Atomic-level description of ubiquitin folding. *Proceedings of the National Academy of Sciences*, 110(15):5915–5920, apr 2013.
- [73] Ignasi Buch, M. J. Harvey, T. Giorgino, D. P. Anderson, and G. De Fabritiis. High-Throughput All-Atom Molecular Dynamics Simulations Using Distributed Computing. *Journal of Chemical Information and Modeling*, 50(3):397–403, mar 2010.

- [74] Nathaniel Stanley, Santiago Esteban-Martín, and Gianni De Fabritiis. Kinetic modulation of a disordered protein domain by phosphorylation. *Nature Communications*, 5(May):5272, oct 2014.
- [75] David E. Shaw, J.P. Grossman, Joseph A. Bank, Brannon Batson, J. Adam Butts, Jack C. Chao, Martin M. Deneroff, Ron O. Dror, Amos Even, Christopher H. Fenton, Anthony Forte, Joseph Gagliardo, Gennette Gill, Brian Greskamp, C. Richard Ho, Douglas J. Ierardi, Lev Iserovich, Jeffrey S. Kuskin, Richard H. Larson, Timothy Layman, Li-Siang Lee, Adam K. Lerer, Chester Li, Daniel Killebrew, Kenneth M. Mackenzie, Shark Yeuk-Hai Mok, Mark A. Moraes, Rolf Mueller, Lawrence J. Nociolo, Jon L. Peticolas, Terry Quan, Daniel Ramot, John K. Salmon, Daniele P. Scarpazza, U. Ben Schafer, Naseer Siddique, Christopher W. Snyder, Jochen Spengler, Ping Tak Peter Tang, Michael Theobald, Horia Toma, Brian Towles, Benjamin Vitale, Stanley C. Wang, and Cliff Young. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. In *SC14: International Conference for High Performance Computing, Networking, Storage and Analysis*, volume 2015-Janua, pages 41–53. IEEE, nov 2014.
- [76] I. Ohmura, G. Morimoto, Y. Ohno, A. Hasegawa, and M. Taiji. MDGRAPE-4: a special-purpose computer system for molecular dynamics simulations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372(2021):20130387–20130387, jun 2014.
- [77] Andrew R Leach. *Molecular modelling: principles and applications*. Pearson education, 2001.
- [78] M.A. González. Force fields and molecular dynamics simulations. *École thématique de la Société Française de la Neutronique*, 12:169–200, jun 2011.
- [79] William C. Swope, Hans C. Andersen, Peter H. Berens, and Kent R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, 76(1):637–649, jan 1982.
- [80] Patrick Walsh, Karen Simonetti, and Simon Sharpe. Core Structure of Amyloid Fibrils Formed by Residues 106126 of the Human Prion Protein. *Structure*, 17(3):417–426, mar 2009.
- [81] Ping Cao, Peter Marek, Harris Noor, Vadim Patsalo, Ling-Hsien Tu, Hui Wang, Andisheh Abedini, and Daniel P. Raleigh. Islet amyloid: From fundamental

- biophysics to mechanisms of cytotoxicity. *FEBS Letters*, 587(8):1106–1118, apr 2013.
- [82] Lauren E Buchanan, Emily B Dunkelberger, Huong Q Tran, Pin-Nan Cheng, Chi-Cheng Chiu, Ping Cao, Daniel P Raleigh, Juan J de Pablo, James S Nowick, and Martin T Zanni. Mechanism of IAPP amyloid fibril formation involves an intermediate with a transient β -sheet. *Proceedings of the National Academy of Sciences of the United States of America*, 110:19285–90, 2013.
- [83] Eugene L Opie. The relation of diabetes mellitus to lesions of the pancreas. hyaline degeneration of the islands of langerhans. *The Journal of experimental medicine*, 5(5):527, 1901.
- [84] P. Westermark, C. Wernstedt, E. Wilander, D. W. Hayden, T. D. O’Brien, and K. H. Johnson. Amyloid fibrils in human insulinoma and islets of Langerhans of the diabetic cat are derived from a neuropeptide-like protein also present in normal islet cells. *Proceedings of the National Academy of Sciences*, 84(11):3881–3885, jun 1987.
- [85] G J Cooper, A C Willis, A Clark, R C Turner, R B Sim, and K B Reid. Purification and characterization of a peptide from amyloid-rich pancreases of type 2 diabetic patients. *Proceedings of the National Academy of Sciences*, 84(23):8628–8632, dec 1987.
- [86] Daniel F. Moriarty and Daniel P. Raleigh. Effects of Sequential Proline Substitutions on Amyloid Formation by Human Amylin 20 - 29 . *Biochemistry*, 38(6):1811–1818, feb 1999.
- [87] Konstantinos Tenidis, Michaela Waldner, Jürgen Bernhagen, Wolfgang Fischle, Michael Bergmann, Marco Weber, Marie-Luise Merkle, Wolfgang Voelter, Herwig Brunner, and Aphrodite Kapurniotu. Identification of a penta- and hexapeptide of islet amyloid polypeptide (IAPP) with amyloidogenic and cytotoxic properties. *Journal of Molecular Biology*, 295(4):1055–1071, jan 2000.
- [88] Jillian Madine, Edward Jack, Peter G. Stockley, Sheena E. Radford, Louise C. Serpell, and David A. Middleton. Structural Insights into the Polymorphism of Amyloid-Like Fibrils Formed by Region 2029 of Amylin Revealed by Solid-State NMR and X-ray Fiber Diffraction. *Journal of the American Chemical Society*, 130(45):14990–15001, nov 2008.

- [89] M Cecchini, R Curcio, M Pappalardo, R Melki, and a Caffisch. A molecular dynamics approach to the structural characterization of amyloid aggregation. *Journal Of Molecular Biology*, 357(4):1306–1321, April 2006.
- [90] Ehud Gazit. Mechanisms of amyloid fibril self-assembly and inhibition. Model short peptides as a key research tool. *The FEBS journal*, 272(23):5971–8, December 2005.
- [91] Rozita Laghaei, Normand Mousseau, and Guanghong Wei. Structure and thermodynamics of amylin dimer studied by Hamiltonian-temperature replica exchange molecular dynamics simulations. *The journal of physical chemistry. B*, 115(12):3146–54, March 2011.
- [92] Bon W. Koo and Andrew D. Miranker. Contribution of the intrinsic disulfide to the assembly mechanism of islet amyloid. *Protein Science*, 14(1):231–239, jan 2009.
- [93] Melania Apostolidou, Sajith A. Jayasinghe, and Ralf Langen. Structure of α -Helical Membrane-bound Human Islet Amyloid Polypeptide and Its Implications for Membrane-mediated Misfolding. *Journal of Biological Chemistry*, 283(25):17205–17210, jun 2008.
- [94] Guizhao Liang, Jun Zhao, Xiang Yu, and Jie Zheng. Comparative molecular dynamics study of human islet amyloid polypeptide (IAPP) and rat IAPP oligomers. *Biochemistry*, 52(6):1089–100, February 2013.
- [95] Mark S Palmer, Aidan J Dryden, J Trevor Hughes, and John Collinge. Homozygous prion protein genotype predisposes to sporadic CreutzfeldtJakob disease. *Nature*, 352(6333):340–342, jul 1991.
- [96] W J Strittmatter, a M Saunders, D Schmechel, M Pericak-Vance, J Enghild, G S Salvesen, and a D Roses. Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proceedings of the National Academy of Sciences of the United States of America*, 90(March):1977–1981, 1993.
- [97] S. Prusiner. Novel proteinaceous infectious particles cause scrapie. *Science*, 216(4542):136–144, April 1982.
- [98] Stanley B. Prusiner, Michael P. McKinley, Karen A. Bowman, David C. Bolton, Paul E. Bendheim, Darlene F. Groth, and George G. Glenner. Scrapie prions

- aggregate to form amyloid-like birefringent rods. *Cell*, 35(2):349–358, December 1983.
- [99] Michael P. McKinley, David C. Bolton, and Stanley B. Prusiner. A protease-resistant protein is a structural component of the Scrapie prion. *Cell*, 35(1):57–62, November 1983.
- [100] S. B. Prusiner. Nobel Lecture: Prions. *Proceedings of the National Academy of Sciences*, 95(23):13363–13383, nov 1998.
- [101] Frédéric Eghiaian, Jeanne Grosclaude, Stéphanie Lesceu, Pascale Debey, Bénédicte Doublet, Eric Tréguer, Human Rezaei, and Marcel Knossow. Insight into the PrPC→PrPSc conversion from the structures of antibody-bound ovine prion scrapie-susceptibility variants. *Proceedings of the National Academy of Sciences of the United States of America*, 101(28):10254–9, jul 2004.
- [102] Masakazu Sekijima, Chie Motono, Satoshi Yamasaki, Kiyotoshi Kaneko, and Yutaka Akiyama. Molecular dynamics simulation of dimeric and monomeric forms of human prion protein: insight into dynamics and properties. *Biophysical journal*, 85(2):1176–85, aug 2003.
- [103] Mari L DeMarco and Valerie Daggett. From conversion to aggregation: protofibril formation of the prion protein. *Proceedings of the National Academy of Sciences of the United States of America*, 101(8):2293–8, feb 2004.
- [104] Jonathan O. Speare, Thomas S. Rush, Marshall E. Bloom, and Byron Caughey. The Role of Helix 1 Aspartates and Salt Bridges in the Stability and Conversion of Prion Protein. *Journal of Biological Chemistry*, 278(14):12522–12529, apr 2003.
- [105] G Forloni, N Angeretti, R Chiesa, E Monzani, M Salmona, O Bugiani, and F Tagliavini. Neurotoxicity of a prion protein fragment. *Nature*, 362(6420):543–6, April 1993.
- [106] Gianluigi Forloni, Roberto Del Bo, Nadia Angeretti, Roberto Chiesa, Simona Smioldo, Rinaldo Doni, Elena Ghibaudi, Mario Salmona, Monica Porro, Laura Verga, Giorgio Giaccone, Orso Bugiani, and Fabrizio Tagliavini. A Neurotoxic Prion Protein Fragment Induces Rat Astroglial Proliferation and Hypertrophy. *European Journal of Neuroscience*, 6(9):1415–1422, September 1994.
- [107] Giuseppe Di Natale, Giuseppe Impellizzeri, and Giuseppe Pappalardo. Conformational properties of peptide fragments homologous to the 106-114 and 106-126

- residues of the human prion protein: a CD and NMR spectroscopic study. *Organic & biomolecular chemistry*, 3(3):490–7, February 2005.
- [108] Yaakov Levy, Eilat Hanan, Beka Solomon, and Oren M. Becker. Helix-coil transition of PrP106-126: Molecular dynamic study. *Proteins: Structure, Function, and Genetics*, 45(4):382–396, dec 2001.
- [109] Alessandra Villa, Alan E Mark, Gloria A A Saracino, Ugo Cosentino, Demetrio Pitea, Giorgio Moro, and Mario Salmona. Conformational polymorphism of the PrP106-126 peptide in different environments: a molecular dynamics study. *The journal of physical chemistry. B*, 110(3):1423–8, January 2006.
- [110] Megan Grabenauer, Chun Wu, Patricia Soto, Joan Emma Shea, and Michael T. Bowers. Oligomers of the prion protein fragment 106-126 are likely assembled from beta-hairpins in solution, and methionine oxidation inhibits assembly without altering the peptide’s monomeric conformation. *Journal of the American Chemical Society*, 132(9):532–539, 2010.
- [111] Lulu Ning, Jingjing Guo, Qifeng Bai, Nengzhi Jin, Huanxiang Liu, and Xiaojun Yao. Structural diversity and initial oligomerization of PrP106-126 studied by replica-exchange and conventional molecular dynamics simulations. *PloS one*, 9(2):e87266, January 2014.
- [112] Jindich Kopeček. Hydrogel biomaterials: A smart future? *Biomaterials*, 28(34):5185–5192, dec 2007.
- [113] Meital Reches and Ehud Gazit. Self-assembly of peptide nanotubes and amyloid-like structures by charged-termini-capped diphenylalanine peptide analogues. *Israel Journal of Chemistry*, 45(3):363–371, jun 2005.
- [114] Assaf Mahler, Meital Reches, Meirav Rechter, Smadar Cohen, and Ehud Gazit. Rigid, Self-Assembled Hydrogel Composed of a Modified Aromatic Dipeptide. *Advanced Materials*, 18(11):1365–1370, jun 2006.
- [115] Izhack Cherny and Ehud Gazit. Amyloids: Not Only Pathological Agents but Also Ordered Nanomaterials. *Angewandte Chemie International Edition*, 47(22):4062–4069, may 2008.
- [116] I. W. Hamley. Self-assembly of amphiphilic peptides. *Soft Matter*, 7(9):4122, 2011.

- [117] S. Ganesh and R. Jayakumar. Structural transitions involved in a novel amyloid-like β -sheet assemblage of tripeptide derivatives. *Biopolymers*, 70(3):336–345, nov 2003.
- [118] Sergey E. Paramonov, Ho-Wook Jun, and Jeffrey D. Hartgerink. Self-assembly of peptide-amphiphile nanofibers: the roles of hydrogen bonding and amphiphilic packing. *Journal of the American Chemical Society*, 128(22):7291–8, jun 2006.
- [119] Sergey E. Paramonov, HoWook Jun, and Jeffrey D. Hartgerink. Modulation of PeptideAmphiphile Nanofibers via Phospholipid Inclusions. *Biomacromolecules*, 7(1):24–26, jan 2006.
- [120] One Sun Lee, Samuel I. Stupp, and George C. Schatz. Atomistic molecular dynamics simulations of peptide amphiphile self-assembly into cylindrical nanofibers. *Journal of the American Chemical Society*, 133(10):3677–3683, 2011.
- [121] E. Deniz Tekin. Molecular dynamics simulations of self-assembled peptide amphiphile based cylindrical nanofibers. *RSC Adv.*, 5(82):66582–66590, 2015.
- [122] Yuri S. Velichko, Samuel I. Stupp, and Monica Olvera de la Cruz. Molecular Simulation Study of Peptide Amphiphile Self-Assembly. *The Journal of Physical Chemistry B*, 112(8):2326–2334, feb 2008.
- [123] Ghibom Bhak, Soonkoo Lee, Jae Woo Park, Sunghyun Cho, and Seung R Paik. Amyloid hydrogel derived from curly protein fibrils of alpha-synuclein. *Biomaterials*, 31(23):5986–95, aug 2010.
- [124] Rohan A. Hule, Radhika P. Nagarkar, Boualem Hammouda, Joel P. Schneider, and Darrin J. Pochan. Dependence of Self-Assembled Peptide Hydrogel Network Structure on Local Fibril Nanostructure. *Macromolecules*, 42(18):7137–7145, sep 2009.
- [125] David Kaplan, W Wade Adams, Barry Farmer, and Christopher Viney. *Silk polymers: materials science and biotechnology*. ACS Publications, 1993.
- [126] Shi-Juang He, Regina Valluzzi, and Samuel P. Gido. Silk I structure in Bombyx mori silk foams. *International Journal of Biological Macromolecules*, 24(2-3):187–195, mar 1999.
- [127] Tsutomu Yamane, Kôsuken Umemura, and Tetsuo Asakura. The Structural Characteristics of Bombyx mori Silk Fibroin before Spinning As Studied with Molecular Dynamics Simulation. *Macromolecules*, 35(23):8831–8838, nov 2002.

- [128] Priyanka Dubey, Sumit Murab, Sandip Karmakar, Pramit K. Chowdhury, and Sourabh Ghosh. Modulation of Self-Assembly Process of Fibroin: An Insight for Regulating the Conformation of Silk Biomaterials. *Biomacromolecules*, 16(12):3936–3944, dec 2015.
- [129] Pim W J M Frederix, Rein V Ulijn, Neil T Hunt, and Tell Tuttle. Virtual Screening for Dipeptide Aggregation: Toward Predictive Tools for Peptide Self-Assembly. *The journal of physical chemistry letters*, 2(19):2380–2384, 2011.
- [130] Pim W J M Frederix, Gary G Scott, Yousef M Abul-Haija, Daniela Kalafatovic, Charalampos G Pappas, Nadeem Javid, Neil T Hunt, Rein V Ulijn, and Tell Tuttle. Exploring the sequence space for (tri-)peptide self-assembly to design and discover new hydrogels. *Nature Chemistry*, 7(1):30–37, dec 2014.
- [131] Siewert J. Marrink, H. Jelger Risselada, Serge Yefimov, D. Peter Tieleman, and Alex H. De Vries. The MARTINI force field: Coarse grained model for biomolecular simulations. *Journal of Physical Chemistry B*, 111(27):7812–7824, 2007.
- [132] Luca Monticelli, Senthil K Kandasamy, Xavier Periole, Ronald G Larson, D Peter Tieleman, and Siewert-Jan Marrink. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *Journal of Chemical Theory and Computation*, 4(5):819–834, may 2008.
- [133] M Halbrügge and U Walter. Purification of a vasodilator-regulated phosphoprotein from human platelets. *European journal of biochemistry*, 185(1):41–50, oct 1989.
- [134] K. Niebuhr. A novel proline-rich motif present in ActA of *Listeria monocytogenes* and cytoskeletal proteins is the ligand for the EVH1 domain, a protein module present in the Ena/VASP family. *The EMBO Journal*, 16(17):5433–5444, sep 1997.
- [135] L J Ball, R Kühne, B Hoffmann, A Häfner, P Schmieder, R Volkmer-Engert, M Hof, M Wahl, J Schneider-Mergener, U Walter, H Oschkinat, and T Jarchau. Dual epitope recognition by the VASP EVH1 domain modulates polyproline ligand specificity and binding affinity. *The EMBO journal*, 19(18):4903–4914, 2000.
- [136] Tingjun Hou, Ken Chen, William a. McLaughlin, Benzhuo Lu, and Wei Wang. Computational analysis and prediction of the binding motif and protein interacting partners of the Abl SH3 domain. *PLoS Computational Biology*, 2(1):0046–0055, 2006.

- [137] Samuel Genheden and Ulf Ryde. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery*, 10(5):449–461, may 2015.
- [138] James C. Gumbart, Benoît Roux, and Christophe Chipot. Standard Binding Free Energies from Computer Simulations: What Is the Best Strategy? *Journal of Chemical Theory and Computation*, 9(1):794–802, jan 2013.
- [139] Noga Gal, Ahiud Morag, Sofiya Kolusheva, Roland Winter, Meytal Landau, and Raz Jelinek. Lipid bilayers significantly modulate cross-fibrillation of two distinct amyloidogenic peptides. *Journal of the American Chemical Society*, 135(36):13582–9, September 2013.
- [140] A. Ott, R. P. Stolk, F. van Harskamp, H. a. P. Pols, A. Hofman, and M. M. B. Breteler. Diabetes mellitus and the risk of dementia: The Rotterdam Study. *Neurology*, 53(9):1937–1937, dec 1999.
- [141] Brian O’Nuallain, Angela D Williams, Per Westermark, and Ronald Wetzel. Seeding Specificity in Amyloid Growth Induced by Heterologous Fibrils. *Journal of Biological Chemistry*, 279(17):17490–17499, apr 2004.
- [142] Khi Pin Chua, Lock Yue Chew, and Yuguang Mu. Replica Exchange Molecular Dynamics Simulation of Cross-Fibrillation of IAPP and PrP106-126. *Proteins: Structure, Function, and Bioinformatics*, (April):1–13, may 2016.
- [143] He Sun, Zhu Qiao, Khi Pin Chua, Alma Tursic, Si Hui Koh, Xu Liu, Yong-Gui Gao, Yuguang Mu, Xingliang Hou, and Yansong Miao. Negative Regulation of Formin-Mediated Actin Assembly by Profilin 3 is Positively Involved in Arabidopsis Innate Immunity.
- [144] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic Local Alignment Search Tool 2Department of Computer Science. *Journal of Molecular Biology*, 215:403–410, 1990.
- [145] M Miyazono, T Kitamoto, T Iwaki, and J Tateishi. Colocalization of prion protein and beta protein in the same amyloid plaques in patients with Gerstmann-Sträussler syndrome. *Acta neuropathologica*, 83:333–339, 1992.
- [146] Jed J W Wiltzius, Stuart A Sievers, Michael R Sawaya, and David Eisenberg. Atomic structures of IAPP (amylin) fusions suggest a mechanism for fibrillation and the role of insulin in the process. *Protein science : a publication of the Protein Society*, 18(7):1521–30, July 2009.

-
- [147] Alireza Roostaei, Sébastien Côté, and Xavier Roucou. Aggregation and amyloid fibril formation induced by chemical dimerization of recombinant prion protein in physiological-like conditions. *The Journal of biological chemistry*, 284(45):30907–16, November 2009.
- [148] Alireza Roostaei, Simon Beaudoin, Antanas Staskevicius, and Xavier Roucou. Aggregation and neurotoxicity of recombinant α -synuclein aggregates initiated by dimerization. *Molecular neurodegeneration*, 8:5, January 2013.
- [149] N Guex and M C Peitsch. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, 18(15):2714–23, December 1997.
- [150] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, mar 2008.
- [151] Berk Hess, Henk Bekker, Herman J. C Berendsen, and Johannes G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472, sep 1997.
- [152] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An $N\log(N)$ method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089, 1993.
- [153] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity-rescaling. *The Journal of Chemical Physics*, 126(1):014101, mar 2008.
- [154] Alexandra Patriksson and David van der Spoel. A temperature predictor for parallel tempering simulations. *Physical chemistry chemical physics : PCCP*, 10(15):2073–7, April 2008.
- [155] Yuguang Mu, Phuong H Nguyen, and Gerhard Stock. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins*, 58(1):45–52, January 2005.
- [156] Rashmi Kumari, Rajendra Kumar, and Andrew Lynn. g_mmpbsa —A GROMACS Tool for High-Throughput MM-PBSA Calculations. *Journal of Chemical Information and Modeling*, 54(7):1951–1962, jul 2014.

- [157] Tingjun Hou, Junmei Wang, Youyong Li, and Wei Wang. Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling*, 51(1):69–82, jan 2011.
- [158] M Cecchini, F Rao, M Seeber, and A Caffisch. Replica exchange molecular dynamics simulations of amyloid peptide aggregation. *The Journal of Chemical Physics*, 121(21):10748, dec 2004.
- [159] Phuong H. Nguyen, Gerhard Stock, Emil Mittag, Chin-Kun Hu, and Mai Suan Li. Free energy landscape and folding mechanism of a β -hairpin in explicit water: A replica exchange molecular dynamics study. *Proteins: Structure, Function, and Bioinformatics*, 61(4):795–808, oct 2005.
- [160] K.Y. Sanbonmatsu and A.E. García. Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. *Proteins: Structure, Function, and Bioinformatics*, 46(2):225–234, feb 2002.
- [161] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, dec 1983.
- [162] Jessica a Williamson and Andrew D Miranker. Direct detection of transient α -helical states in islet amyloid polypeptide. *Protein Science*, 16(1):110–117, jan 2007.
- [163] Ricardo Simão Vieira-Pires and João Henrique Morais-Cabral. 310 helices in channels and other membrane proteins. *The Journal of General Physiology*, 136(6):585–592, nov 2010.
- [164] Michael F. Jobling, Leanne R. Stewart, Anthony R. White, Catriona McLean, Anna Friedhuber, Fran Maher, Konrad Beyreuther, Colin L. Masters, Colin J. Barrow, Steven J. Collins, and Roberto Cappai. The hydrophobic core sequence modulates the neurotoxic and secondary structure properties of the prion peptide 106-126. *Journal of neurochemistry*, 73(4):1557–65, oct 1999.
- [165] Lydia Young, Hlengisizwe Ndlovu, Tom W. Knapman, Sarah A. Harris, Sheena E. Radford, and Alison E. Ashcroft. Monitoring oligomer formation from self-aggregating amylin peptides using ESI-IMS-MS. *International Journal for Ion Mobility Spectrometry*, 16(1):29–39, mar 2013.

- [166] Michael R. Shirts and John D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of Chemical Physics*, 129(12):124105, 2008.
- [167] John D. Chodera, William C. Swope, Frank Noé, Jan-Hendrik Prinz, Michael R. Shirts, and Vijay S. Pande. Dynamical reweighting: Improved estimates of dynamical properties from simulations at multiple temperatures. *The Journal of Chemical Physics*, 134(24):244107, jun 2011.
- [168] 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: Structure, Function, and Bioinformatics*, 78(8):NA–NA, jun 2010.
- [169] Drazen Petrov and Bojan Zagrovic. Are Current Atomistic Force Fields Accurate Enough to Study Proteins in Crowded Environments? *PLoS Computational Biology*, 10(5):e1003638, may 2014.
- [170] Ravi Prakash Reddy Nanga, Jeffrey R Brender, Jiadi Xu, Gianluigi Veglia, and Ayyalusamy Ramamoorthy. Structures of rat and human islet amyloid polypeptide IAPP(1-19) in micelles by NMR spectroscopy. *Biochemistry*, 47(48):12689–97, December 2008.
- [171] Maarten F M Engel, HaciAli Yigittop, Ronald C Elgersma, Dirk T S Rijkers, Rob M J Liskamp, Ben de Kruijff, Jo W M Höppener, and J Antoinette Killian. Islet amyloid polypeptide inserts into phospholipid monolayers as monomer. *Journal of molecular biology*, 356(3):783–9, February 2006.
- [172] D H J Lopes, A Meister, A Gohlke, A Hauser, A Blume, and R Winter. Mechanism of islet amyloid polypeptide fibrillation at lipid interfaces studied by infrared reflection absorption spectroscopy. *Biophysical journal*, 93(9):3132–41, November 2007.
- [173] 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 & dynamics*, 32(November 2014):1651–69, 2014.
- [174] Robert B. Best, Nicolae-Viorel Buchete, and Gerhard Hummer. Are Current Molecular Dynamics Force Fields too Helical? *Biophysical Journal*, 95(1):L07–L09, jul 2008.

- [175] Peter L. Freddolino, Sanghyun Park, Benoît Roux, and Klaus Schulten. Force Field Bias in Protein Folding Simulations. *Biophysical Journal*, 96(9):3772–3780, may 2009.
- [176] Jiri Vymetal and Jiri Vondrasek. Critical assessment of current force fields. Short peptide test case. *Journal of Chemical Theory and Computation*, 9:441–451, 2013.
- [177] Nicholas F. Dupuis, Chun Wu, Joan Emma Shea, and Michael T. Bowers. The amyloid formation mechanism in human IAPP: Dimers have ??-strand monomer-monomer interfaces. *Journal of the American Chemical Society*, 133:7240–7243, 2011.
- [178] Erika Andreetto, Li-Mei Yan, Marianna Tatarek-Nossol, Aleksandra Velkova, Ronald Frank, and Aphrodite Kapurniotu. Identification of Hot Regions of the A β -IAPP Interaction Interface as High-Affinity Binding Sites in both Cross- and Self-Association. *Angewandte Chemie International Edition*, 49(17):3081–3085, mar 2010.
- [179] Yufen Zhang, Xian Zhao, and Yuguang Mu. Conformational transition map of an RNA GCAA tetraloop explored by replica-exchange molecular dynamics simulation. *Journal of Chemical Theory and Computation*, 5:1146–1154, 2009.
- [180] Boon Chong Goh, Hon Wai Leong, Xiaohui Qu, and Lock Yue Chew. The mechanism of antiparallel β -sheet formation based on conditioned self-avoiding walk. *The European physical journal. E, Soft matter*, 35(4):9704, April 2012.
- [181] Rozita Laghaei, Normand Mousseau, and Guanghong Wei. Effect of the disulfide bond on the monomeric structure of human amylin studied by combined Hamiltonian and temperature replica exchange molecular dynamics simulations. *The journal of physical chemistry. B*, 114(20):7071–7, May 2010.
- [182] Paul A. Guerette, Shawn Hoon, Dawei Ding, Shahrouz Amini, Admir Masic, Vyidianathan Ravi, Byrappa Venkatesh, James C. Weaver, and Ali Miserez. Nanoconfined β -Sheets Mechanically Reinforce the Supra-Biomolecular Network of Robust Squid Sucker Ring Teeth. *ACS Nano*, 8(7):7170–7179, jul 2014.
- [183] Shu Hui Hiew, Paul A. Guerette, Ondrej J. Zvarec, Margaret Phillips, Feng Zhou, Haibin Su, Konstantin Pervushin, Brendan P. Orner, and Ali Miserez. Modular peptides from the thermoplastic squid sucker ring teeth form amyloid-like cross- β supramolecular networks. *Acta Biomaterialia*, 46:41–54, Dec 2016.

- [184] Charles J. Bowerman, Derek M. Ryan, David A. Nissan, and Bradley L. Nilsson. The effect of increasing hydrophobicity on the self-assembly of amphipathic β -sheet peptides. *Molecular BioSystems*, 5(9):1058, 2009.
- [185] Leonid Breydo, Amanda E. Sales, Telma Frege, Mark C. Howell, Boris Y. Zaslavsky, and Vladimir N. Uversky. Effects of Polymer Hydrophobicity on Protein Structure and Aggregation Kinetics in Crowded Milieu. *Biochemistry*, 54(19):2957–2966, may 2015.
- [186] Semen O. Yesylevskyy, Lars V. Schäfer, Durba Sengupta, and Siewert J. Marrink. Polarizable Water Model for the Coarse-Grained MARTINI Force Field. *PLoS Computational Biology*, 6(6):e1000810, jun 2010.
- [187] Siewert J. Marrink, Alex H. de Vries, and Alan E. Mark. Coarse Grained Model for Semiquantitative Lipid Simulations. *The Journal of Physical Chemistry B*, 108(2):750–760, jan 2004.
- [188] Riccardo Baron, Daniel Trzesniak, Alex H. de Vries, Andreas Elsener, Siewert J. Marrink, and Wilfred F. van Gunsteren. Comparison of Thermodynamic Properties of Coarse-Grained and Atomic-Level Simulation Models. *ChemPhysChem*, 8(3):452–461, feb 2007.
- [189] Junichi Higo, Jinzen Ikebe, Narutoshi Kamiya, and Haruki Nakamura. Enhanced and effective conformational sampling of protein molecular systems for their free energy landscapes. *Biophysical Reviews*, 4:27–44, 2012.
- [190] Cameron Abrams and Giovanni Bussi. Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration. *Entropy*, 16:163–199, 2013.
- [191] Omar Valsson, Pratyush Tiwary, and Michele Parrinello. Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annual Review of Physical Chemistry*, 67(1):annurev-physchem-040215-112229, 2016.
- [192] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Physical Review Letters*, 100(2):020603, jan 2008.
- [193] James F. Dama, Michele Parrinello, and Gregory A. Voth. Well-Tempered Metadynamics Converges Asymptotically. *Physical Review Letters*, 112(24):240602, jun 2014.

- [194] Yunqiang Bian, Jian Zhang, Jun Wang, and Wei Wang. On the accuracy of metadynamics and its variations in a protein folding process. *Molecular Simulation*, 41(9):752–763, jun 2015.
- [195] Fabio Pietrucci and Alessandro Laio. A Collective Variable for the Efficient Exploration of Protein Beta-Sheet Structures: Application to SH3 and GB1. *Journal of Chemical Theory and Computation*, 5(9):2197–2201, sep 2009.
- [196] Stefano Piana and Alessandro Laio. A Bias-Exchange Approach to Protein Folding. *The Journal of Physical Chemistry B*, 111(17):4553–4559, may 2007.
- [197] Gül H. Zerze, Cayla M. Miller, Daniele Granata, and Jeetain Mittal. Free Energy Surface of an Intrinsically Disordered Protein: Comparison between Temperature Replica Exchange Molecular Dynamics and Bias-Exchange Metadynamics. *Journal of Chemical Theory and Computation*, 11(6):2776–2782, jun 2015.
- [198] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E. Mark, and Herman J. C. Berendsen. GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry*, 26(16):1701–1718, dec 2005.
- [199] Gareth A. Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. PLUMED 2: New feathers for an old bird. *Computer Physics Communications*, 185(2):604–613, feb 2014.
- [200] 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*, 78(8):1950–8, jun 2010.
- [201] Timothy J. Richmond. Solvent accessible surface area and excluded volume in proteins. *Journal of Molecular Biology*, 178(1):63–89, sep 1984.
- [202] Xevi Biarnés, Fabio Pietrucci, Fabrizio Marinelli, and Alessandro Laio. METAGUI. A VMD interface for analyzing metadynamics and molecular dynamics simulations. *Computer Physics Communications*, 183(1):203–211, jan 2012.
- [203] Fabrizio Marinelli, Fabio Pietrucci, Alessandro Laio, and Stefano Piana. A kinetic model of Trp-cage folding from multiple biased molecular dynamics simulations. *PLoS Computational Biology*, 5(8), 2009.
- [204] Richa Singh, Rohit Bansal, Anurag Singh Rathore, and Gaurav Goel. Equilibrium Ensembles for Insulin Folding from Bias-Exchange Metadynamics. *Biophysical Journal*, 112(8):1571–1585, apr 2017.

- [205] Akira Matsumoto, Jingsong Chen, Adam L Collette, Ung-jin Kim, Gregory H Altman, Peggy Cebe, and David L Kaplan. Mechanisms of Silk Fibroin Sol?Gel Transitions. *The Journal of Physical Chemistry B*, 110(43):21630–21638, nov 2006.
- [206] Sukumaran Santhosh Babu, Vakayil K. Praveen, and Ayyappanpillai Ajayaghosh. Functional π -Gelators and Their Applications. *Chemical Reviews*, 114(4):1973–2129, feb 2014.
- [207] Benjamin P. Partlow, Mehran Bagheri, James L. Harden, and David L. Kaplan. Tyrosine Templating in the Self-Assembly and Crystallization of Silk Fibroin. *Biomacromolecules*, 17(11):3570–3579, nov 2016.
- [208] Naresh Thota, Zhonglin Luo, Zhongqiao Hu, and Jianwen Jiang. Self-Assembly of Amphiphilic Peptide (AF) 6 H 5 K 15 : Coarse-Grained Molecular Dynamics Simulation. *The Journal of Physical Chemistry B*, 117(33):9690–9698, aug 2013.
- [209] Heng Zhang, Hua Wang, Guiying Xu, and Shiling Yuan. A molecular dynamics simulation of N-(fluorenyl-9-methoxycarbonyl)-dipeptides supramolecular hydrogel. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 417:217–223, 2013.
- [210] Cong Guo, Yin Luo, Ruhong Zhou, and Guanghong Wei. Triphenylalanine peptides self-assemble into nanospheres and nanorods that are different from the nanovesicles and nanotubes formed by diphenylalanine peptides. *Nanoscale*, 6(5):2800–11, 2014.
- [211] Xiao Ding Xu, Chang Sheng Chen, Bo Lu, Si Xue Cheng, Xian Zheng Zhang, and Ren Xi Zhuo. Coassembly of oppositely charged short peptides into Well-Defined supramolecular hydrogels. *Journal of Physical Chemistry B*, 114(7):2365–2372, 2010.
- [212] Harvey Lodish, Arnold Berk, S Lawrence Zipursky, Paul Matsudaira, David Baltimore, and James Darnell. The actin cytoskeleton. 2000.
- [213] Roberto Dominguez and Kenneth C. Holmes. Actin Structure and Function. *Annual Review of Biophysics*, 40(1):169–186, jun 2011.
- [214] Wilma A. Hofmann. Cell and molecular biology of nuclear actin. *International review of cell and molecular biology*, 273:219–63, 2009.

-
- [215] Toshiro Oda, Heiko Stegmann, Rasmus R Schröder, Keiichi Namba, and Yuichiro Maéda. Modeling of the F-actin structure. *Advances in experimental medicine and biology*, 592:385–401, 2007.
- [216] K C Holmes, D Popp, W Gebhard, and W Kabsch. Atomic model of the actin filament. *Nature*, 347(6288):44–9, sep 1990.
- [217] E H Egelman. The structure of F-actin. *Journal of muscle research and cell motility*, 6(2):129–51, apr 1985.
- [218] R D Mullins, J a Heuser, and T D Pollard. The interaction of Arp2/3 complex with actin: Nucleation, high affinity pointed end capping, and formation of branching networks of filaments. *Proceedings of the National Academy of Sciences*, 95(11):6181–6186, may 1998.
- [219] Isabelle Sagot, Avital A Rodal, James Moseley, Bruce L Goode, and David Pellman. An actin nucleation mechanism mediated by Bni1 and profilin. *Nature cell biology*, 4(8):626–31, 2002.
- [220] Stéphane Romero, Christophe Le Clainche, Dominique Didry, Coumaran Egile, Dominique Pantaloni, and Marie-France Carlier. Formin Is a Processive Motor that Requires Profilin to Accelerate Actin Assembly and Associated ATP Hydrolysis. *Cell*, 119(3):419–429, oct 2004.
- [221] Aditya S Paul, Aditya Paul, Thomas D Pollard, and Thomas Pollard. The role of the FH1 domain and profilin in formin-mediated actin-filament elongation and nucleation. *Current biology : CB*, 18(1):9–19, jan 2008.
- [222] Nicole M Mahoney, Denise A Rozwarski, E Fedorov, Alexander A Fedorov, and Steven C Almo. Profilin binds proline-rich ligands in two distinct amide backbone orientations. *Nature structural biology*, 6(7):666–71, jul 1999.
- [223] Martin Pring, Marie Evangelista, Charles Boone, Changsong Yang, and Sally H. Zigmond. Mechanism of Formin-Induced Nucleation of Actin Filaments. *Biochemistry*, 42(2):486–496, jan 2003.
- [224] Henry N. Higgs. Formin proteins: a domain-based approach. *Trends in Biochemical Sciences*, 30(6):342–353, jun 2005.
- [225] Petri Kursula, Inari Kursula, Marzia Massimi, Young Hwa Song, Joshua Downer, Will a. Stanley, Walter Witke, and Matthias Wilmanns. High-resolution Structural Analysis of Mammalian Profilin 2a Complex Formation with Two Physio-

- logical Ligands: The Formin Homology 1 Domain of mDia1 and the Proline-rich Domain of VASP. *Journal of Molecular Biology*, 375(1):270–290, 2008.
- [226] A A Fedorov, K A Magnus, M H Graupe, E E Lattman, T D Pollard, and S C Almo. X-ray structures of isoforms of the actin-binding protein profilin that differ in their affinity for phosphatidylinositol phosphates. *Proceedings of the National Academy of Sciences of the United States of America*, 91(18):8636–40, aug 1994.
- [227] A A Fedorov, T Ball, N M Mahoney, R Valenta, and S C Almo. The molecular basis for allergen cross-reactivity: crystal structure and IgE-epitope mapping of birch pollen profilin. *Structure (London, England : 1993)*, 5(1):33–45, jan 1997.
- [228] K S Thorn, H E Christensen, R Shigeta, D Huddler, L Shalaby, U Lindberg, N H Chua, and C E Schutt. The crystal structure of a major allergen from plants. *Structure (London, England : 1993)*, 5(1):19–32, jan 1997.
- [229] C E Schutt, J C Myslik, M D Rozycki, N C Goonesekere, and U Lindberg. The structure of crystalline profilin-beta-actin. *Nature*, 365(6449):810–6, oct 1993.
- [230] P. W. Gunning, U. Ghoshdastider, S. Whitaker, D. Popp, and R. C. Robinson. The evolution of compositionally and functionally distinct actin filaments. *Journal of Cell Science*, 128(11):2009–2019, jun 2015.
- [231] Alexei a. Adzhubei, Michael J.E. Sternberg, and Alexander a. Makarov. Polyproline-II Helix in Proteins: Structure and Function. *Journal of Molecular Biology*, 425(12):2100–2132, jun 2013.
- [232] Mahmoud Moradi, Volodymyr Babin, Christopher Roland, and Celeste Sagui. A classical molecular dynamics investigation of the free energy and structure of short polyproline conformers. *The Journal of Chemical Physics*, 133(12):125104, 2010.
- [233] M A Kelly, B W Chellgren, A L Rucker, J M Troutman, M G Fried, A F Miller, and T P Creamer. Host-guest study of left-handed polyproline II helix formation. *Biochemistry*, 40(48):14376–83, dec 2001.
- [234] Adam L. Rucker, Cara T. Pager, Margaret N. Campbell, Joseph E. Qualls, and Trevor P. Creamer. Host-guest scale of left-handed polyproline II helix formation. *Proteins*, 53(1):68–75, oct 2003.

- [235] M A Martí-Renom, A C Stuart, A Fiser, R Sánchez, F Melo, and A Sali. Comparative protein structure modeling of genes and genomes. *Annual review of biophysics and biomolecular structure*, 29:291–325, 2000.
- [236] Maria J. Macias, Silke Wiesner, and Marius Sudol. WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. *FEBS Letters*, 513(1):30–37, feb 2002.
- [237] Marco Biasini, Stefan Bienert, Andrew Waterhouse, Konstantin Arnold, Gabriel Studer, Tobias Schmidt, Florian Kiefer, Tiziano Gallo Cassarino, Martino Bertoni, Lorenza Bordoli, and Torsten Schwede. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, 42(W1):W252–W258, jul 2014.
- [238] David T Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, 292(2):195–202, sep 1999.
- [239] Daniel W. A. Buchan, Federico Minneci, Tim C. O. Nugent, Kevin Bryson, and David T. Jones. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic acids research*, 41(Web Server issue):W349–57, jul 2013.
- [240] H.J.C. Berendsen, D. van der Spoel, and R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1-3):43–56, sep 1995.
- [241] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, 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, sep 2015.
- [242] Jing Huang and Alexander D. MacKerell. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry*, 34(25):2135–2145, sep 2013.
- [243] Federico Fogolari, Alessandro Brigo, and Henriette Molinari. Protocol for MM/PBSA molecular dynamics simulations of proteins. *Biophysical journal*, 85(1):159–166, 2003.
- [244] N A Baker, D Sept, S Joseph, M J Holst, and J A McCammon. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences of the United States of America*, 98(18):10037–41, aug 2001.

- [245] 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(1-2):236–240, jan 1999.
- [246] W A Lim, F M Richards, and R O Fox. Structural determinants of peptide-binding orientation and of sequence specificity in SH3 domains. *Nature*, 372(6504):375–9, nov 1994.
- [247] Muthuraman Meiyappan, Gabriel Birrane, and John A.A. Ladas. Structural Basis for Polyproline Recognition by the FE65 WW Domain. *Journal of Molecular Biology*, 372(4):970–980, sep 2007.
- [248] M.Teresa Pisabarro, Luis Serrano, and Matthias Wilmanns. Crystal structure of the abl-SH3 domain complexed with a designed high-affinity peptide ligand: implications for SH3-ligand interactions. *Journal of Molecular Biology*, 281(3):513–521, aug 1998.
- [249] Neal J Zondlo. Aromatic-Proline Interactions: Electronically Tunable CH/ π Interactions. *Accounts of Chemical Research*, 46(4):1039–1049, apr 2013.
- [250] M Salanoubat, K Lemcke, M Rieger, W Ansorge, M Unseld, B Fartmann, G Valle, H Blöcker, M Perez-Alonso, B Obermaier, M Delseny, M Boutry, L A Grivell, R Mache, P Puigdomènech, V De Simone, N Choisine, F Artigue-nave, C Robert, P Brottier, P Wincker, L Cattolico, J Weissenbach, W Saurin, F Quétier, M Schäfer, S Müller-Auer, C Gabel, M Fuchs, V Benes, E Wurm-bach, H Drzonek, H Erfle, N Jordan, S Bangert, R Wiedelmann, H Kranz, H Voss, R Holland, P Brandt, G Nyakatura, A Vezzi, M D’Angelo, A Pallavicini, S Toppo, B Simionati, A Conrad, K Hornischer, G Kauer, T H Löhnert, G Nord-siek, J Reichelt, M Scharfe, O Schön, M Bargues, J Terol, J Climent, P Navarro, C Collado, A Perez-Perez, B Ottenwälder, D Duchemin, R Cooke, M Laudie, C Berger-Llauro, B Purnelle, D Masuy, M de Haan, A C Maarse, J P Alcaraz, A Cottet, E Casacuberta, A Monfort, A Argiriou, M Flores, R Liguori, D Vi-tale, G Mannhaupt, D Haase, H Schoof, S Rudd, P Zaccaria, H W Mewes, K F Mayer, S Kaul, C D Town, H L Koo, L J Tallon, J Jenkins, T Rooney, M Rizzo, A Walts, T Utterback, C Y Fujii, T P Shea, T H Creasy, B Haas, R Maiti, D Wu, J Peterson, S Van Aken, G Pai, J Militscher, P Sellers, J E Gill, T V Feldblyum, D Preuss, X Lin, W C Nierman, S L Salzberg, O White, J C Venter, C M Fraser, T Kaneko, Y Nakamura, S Sato, T Kato, E Asamizu, S Sasamoto, T Kimura, K Idesawa, K Kawashima, Y Kishida, C Kiyokawa, M Kohara, M Matsumoto, A Matsuno, A Muraki, S Nakayama, N Nakazaki,

- S Shinpo, C Takeuchi, T Wada, A Watanabe, M Yamada, M Yasuda, S Tabata, European Union Chromosome 3 Arabidopsis Sequencing Consortium, Institute for Genomic Research, and Kazusa DNA Research Institute. Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. *Nature*, 408(6814):820–2, dec 2000.
- [251] E. M. Marcotte. Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. *Science*, 285(5428):751–753, jul 1999.
- [252] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering, and L. J. Jensen. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41(D1):D808–D815, jan 2013.
- [253] Ulrich Stelzl, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H. Brembeck, Heike Goehler, Martin Stroedicke, Martina Zenkner, Anke Schoenherr, Susanne Koeppen, Jan Timm, Sascha Mintzlaff, Claudia Abraham, Nicole Bock, Silvia Kietzmann, Astrid Goedde, Engin Toksöz, Anja Droege, Sylvia Krobitsch, Bernhard Korn, Walter Birchmeier, Hans Lehrach, and Erich E. Wanker. A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. *Cell*, 122(6):957–968, sep 2005.
- [254] Christian von Mering, Roland Krause, Berend Snel, Michael Cornell, Stephen G. Oliver, Stanley Fields, and Peer Bork. Comparative assessment of large-scale data sets of proteinprotein interactions. *Nature*, pages 399–403, may 2002.