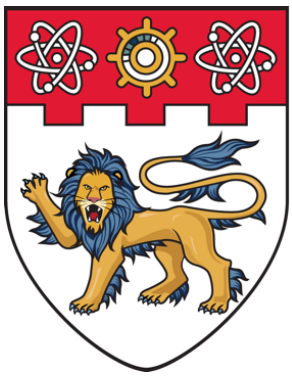

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

APPLICATIONS OF MOLECULAR DYNAMICS SIMULATION IN BIOMOLECULAR FOLDING AND RECOGNITION

NG TZE YANG JUSTIN

SCHOOL OF BIOLOGICAL SCIENCES

2020

APPLICATIONS OF MOLECULAR DYNAMICS SIMULATION IN BIOMOLECULAR FOLDING AND RECOGNITION

NG TZE YANG JUSTIN

SCHOOL OF BIOLOGICAL SCIENCES

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

2020

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research done by me except where otherwise stated in this thesis. The thesis work has not been submitted for a degree or professional qualification to any other university or institution. I declare that this thesis is written by myself and is free of plagiarism and of sufficient grammatical clarity to be examined. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

13032020

. . . . . .
Date

. . . . . .
Ng Tze Yang Justin

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it of sufficient grammatical clarity to be examined. To the best of my knowledge, the thesis is free of plagiarism and the research and writing are those of the candidate's except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

13032020

. . . . . .

Date

. . . . . .

A/Prof Mu Yuguang

Authorship Attribution Statement

(B) This thesis contains material from 2 papers published in the following peer-reviewed journals / from papers accepted at conferences in which I am listed as an author.

Chapter 4 is published as:

Kannaian, B. , Sharma, B. , Phillips, M. , Chowdhury, A. , Manimekalai, M.S.S. , Adav, S.S. , Ng, J.T.Y. , Kumar, A. , Lim, S. , Mu, Y. , Sze, S.K. , Grüber, G. and Pervushin, K. (2019). Abundant neuroprotective chaperone Lipocalin-type prostaglandin D synthase (L-PGDS) disassembles the Amyloid-β fibrils. Sci Rep. 2019. 9(1):12579

The contributions of the co-authors are as follows:

Kannaian, B. prepared samples and performed all biophysical and NMR experiments and wrote the manuscript. Sharma, B prepared conjugates and Phillips, M. captured TEM images. Chowdhury, A. helped in preliminary biophysical studies. Manimekalai, M.S.S. and Grüber, G. performed and analyzed SAXS experiments. Adav, S.S., Kannaian, B. and Sze, S.K. prepared samples for proteomics and analyzed the data. Ng, J.T.Y. and Mu,Y. performed the molecular dynamic simulations. , Kumar, A. and Lim, S. provided ferritin samples. Pervushin, K. devised the experimental plan and supervised the study. Kannaian, B., Sharma, B, Phillips, M., Manimekalai, M.S.S., Adav, S.S., Ng, J.T.Y., Grüber, G. and Pervushin, K. wrote and edited the manuscript. All authors reviewed, revised and approved the manuscript.

Chapter 5 is published as:

Qiao, Z. , Sun, H. , Ng, J.T.Y. , Ma, Q. , Koh, S.H. , Mu, Y. , Miao, Y. and Gao, Y.G. (2019). Structural and computational examination of the Arabidopsis profilin-Poly-P complex reveals mechanistic details in profilin-regulated actin assembly. J Biol Chem. 294(49):18650-18661.

The contributions of the co-authors are as follows:

Qiao, Z. performed crystallography experiments. Sun, H. performed biochemistry experiments. Ng, J.T.Y. performed molecular modelling and molecular dynamics simulations. Ma, Q. and Koh, S.H. performed yeast experiments. Mu, Y. , Miao, Y. and Gao, Y.G. designed the research and supervised the study. All authors wrote, reviewed, revised and approved the manuscript.

13032020

. . . . . .
Date

. . . . . .
Ng Tze Yang Justin

Acknowledgements

I would like to thank to my supervisor, A/Prof. Mu Yuguang, for his constant support, encouragement and guidance during the four years of my Ph.D study. I would also like to thank my co-supervisor A/Prof Julien Lescar for introducing me to the field of Structural Biology.

I am grateful towards my Thesis Advisory Committee members, A/Prof Lu Lanyuan and A/Prof Liu Chuan Fa for their suggestions and comments about research direction during the thesis advisory committee meetings.

I would like to express my gratitude towards my experimental collaborators, Dr. James Ho, Prof. Catharina Svanborg (Lund University, Sweden) (HAMLET project); Ms. Kannaian Bhuvaneswari, Dr. Margaret Phillips, A/Prof Konstantin Pervushin, A/Prof Newman Sze, Prof. Gerhard Gruber (L-PGDS project); Mr. Sun He, Mr. Qiao Zhu, A/Prof Gao Yonggui, and Asst Prof Miao Yansong (Profilin3 Project) for their insightful discussions on the incorporation of experimental data towards molecular modelling.

I would like to especially acknowledge my laboratory mates and seniors, Dr. Liu Yang, Dr. Zheng Liangzhen, Dr. Chua Khi Pin, Dr. Amr Ali Mokhtar Alhossary, Dr. Guo JingJing, Dr. Fan Jingrong, Dr. Ning Lulu, Dr. Yaw Awuni, Dr. Luo Di, Dr. Zhu Guanhua, Dr Zhang Haiping and Mr. Saxena Shikhar for their constant sharing of their research experiences.

Finally, I am also thankful to my family and friends for their constant concern, encouragement, love and support.

Table of Contents

Acknowledgements i
Table of Contents iii
List of Figures vi
List of Tables xi
List of Abbreviations xii
Abstract xiv
1. Introduction 1
1.1 Introduction to Proteins 1
1.2 Introduction to Protein Folding 3
1.3 Biomolecular recognition 5
1.4 Molecular dynamics simulations for studies of biomolecular folding and recognition 8
1.5 Aims and scope of the thesis 10
2. Method 13
2.1 Introduction to Molecular Modelling 13
2.2 Force Field 14
2.3 Classical Molecular Dynamics simulations 15
2.4 Hamiltonian Replica Exchange Molecular Dynamics 17
2.5 Binding Free Energy Calculations 21
2.6 Identification of biomolecular interactions 22
3. In-silico modeling of the Human Alpha-lactalbumin Made Lethal to Tumor cells (HAMLET) Alpha1 complex 27
3.1 Introduction 27
3.1.1 HAMLET (Human Alpha-lactalbumin Made Lethal to Tumor cells) is a tumoricidal complex of Alpha-lactalbumin and oleic acid 27
3.1.2 Biological Effects of HAMLET 29
3.1.3 Structural studies of the HAMLET complex 30
3.1.4 Research Aim and direction 31
3.2 Method 32
3.2.1 Initial model construction 32
3.2.2 Hamiltonian replica exchange molecular dynamics 33
3.2.3 Analysis of simulation data 34
3.3 Results 35

3.3.1 Assessment of simulation convergence 35
3.3.2 Secondary structure analysis reveals higher percentage of helical conformations for the Alpha1-oleate containing system 37
3.3.3 Free energy landscape as a function of Principal components of dihedral angles reveal conformational differences between both systems 40
3.3.4 Geometric clustering of conformations for both systems confirms findings of Dihedral Angle Principal Component Analysis 43
3.3.5 Contact probability calculations between Alpha1 and oleate molecules reveal high contact probabilities between hydrophobic residues and oleate molecules 45
3.3.6 Contact Probability analysis between Alpha1 peptide and Oleate in folded state 46
3.3.7 Perspective of simulation findings 49
3.4. Conclusions and future directions 50
4. Atomistic study of binding between Lipocalin type Prostaglandin D synthase (L-PGDS) and Amyloid Beta peptide(1-40) Aβ(1-40)- An in-silico study 52
4.1 Introduction 52
4.1.1 Alzheimer's Disease is a "Protein Misfolding disease" 52
4.1.2 Lipocalin type Prostaglandin D synthase (L-PGDS) is a major amyloid beta (Aβ) chaperone 53
4.1.3 Aim of project and scope 54
4.2 Method 57
4.2.1 Initial construction of L-PGDS- Aβ40 model 57
4.2.2 Hamiltonian Replica Exchange Molecular Dynamics 58
4.2.3 Contact Probability based clustering method 58
4.2.4 Classical Molecular Dynamics simulations and Analysis 60
4.3 Results and discussion 61
4.3.1 Convergence of H-REMD simulations 61
4.3.2 Contact probability calculations reveal high contact probability residue pairs between L-PGDS and Aβ40 63
4.3.3 Re-labelling of simulation frames with high contact probability residue pairs identifies 3 binding modes of L-PGDS towards Aβ40 66
4.3.4 Biomolecular interactions between L-PGDS and Aβ40 identified through classical molecular dynamics simulations 68
4.3.5 Further Root Mean Square Deviation calculations and Binding Free energy calculations confirm the stability of Binding Mode 1 71
4.3.6 Agreement of L-PGDS and Aβ40 binding model with experimental data 73
4.3.7 Comparison of L-PGDS- Aβ complex and Anticalin US7- Aβ complex 79
4.3.8 Perspective of simulation findings 80
4.4 Conclusion and future direction 81

5. The plasticity of *Arabidopsis thaliana* Profilin3 N-Terminal extension and its role in Formin Poly-proline binding 83
5.1 Introduction 83
5.1.1 Profilins are proteins involved in plant actin cytoskeleton regulation 83
5.1.2 Arabidopsis thaliana Profilin 3 is a unique profilin isoform containing N-Terminus extension 83
5.1.3 Experimental studies of interactions between Profilins and PolyP 85
5.1.4 Aim of Project and scope 88
5.2 Method 90
5.2.1 Hamiltonian Replica Exchange Molecular Dynamics to sample to conformation of AtPRF3 NTE 90
5.2.2 Free Energy Surface to describe conformational Plasticity of AtPRF3 NTE 91
5.2.3 Model construction of AtPRF3Δ22- PolyP complex 93
5.2.4 Classical Molecular Dynamics simulations and binding mode analysis of PolyP bound AtPRF3Δ22 94
5.2.5 Binding Energy analysis 95
5.3 Results and Discussion 96
5.3.1 Plasticity of the AtPRF3 NTE revealed by H-REMD 96
5.3.2 Understanding the interactions in the AtPRF3Δ37-PolyP complex 100
5.3.3 Understanding the interactions between PRF3Δ22 and PolyP-The binding of AtPRF3 NTE to PolyP may involve multiple, adaptive binding modes 102
5.3.4 Alternative model of AtPRF3Δ22SM-PolyP complex with NTE modelled in partial helical conformation 106
5.3.5 Perspective of Simulation Findings 109
5.4 Conclusion and future directions 110
6. Conclusions 111
List of manuscripts in submission/preparation 113
References 114

List of Figures

Figure 1.1. Schematic overview of an amino acid and the amino acid sequence of a short peptide. A. Schematic overview of an amino acid, showing the alpha carbon, amino and carboxyl groups, and the side chain. At pH7, the termini of amino acids are ionized and charged. B. Schematic overview of the amino acid sequence of a short peptide, showing an example consisting of 3 different amino acids (side chains labelled "R1", "R2" and "R3"). The peptide bond is labelled. The N- and C- terminals are indicated, respectively. Figure adapted from reference [2]. (Pg. 2)

Figure 3.1 Introduction to the formation of Human Alpha-lactalbumin Made Lethal to Tumor cells (HAMLET) complex. Human Alpha-lactalbumin binds with oleic acid in the ratio of 1:4-7 and partially unfolds with the removal of Calcium to form the HAMLET complex. Figure adapted from reference [19]. (Pg. 28)

Figure 3.2 Figure summarizing the biological effects of Human Alpha-lactalbumin Made Lethal to Tumor cells (HAMLET). The Biological effects of HAMLET may be described as a broad-spectrum apoptosis like response towards tumor cells without the affecting of healthy, differentiated cells Figure adapted from reference [19]. (Pg.29)

Figure 3.3. Evaluation of Hamiltonian Replica Exchange Molecular Dynamics simulation convergence. Diffusiveness of representative H-REMD replicas from a representative low (effective temperature= 300K, Black), medium (effective temperature= 477.4K, Red) and high (effective temperature= 800K, Blue) effective temperature replica for Alpha1-oleate containing system (A) and Alpha1-apo system (E). Simulation properties as a function of time were plotted. i to i+4 Hydrogen bond number for Alpha1-oleate containing system (B) and Alpha1-apo system (F). Alpha1 peptide Radius of Gyration for the Alpha1-oleate containing system (C) and Alpha1-apo system (G). Number of heavy atom contacts between the Alpha1 peptide and oleate molecules (D) in Alpha1-oleate containing system. For clarity, properties relating to the Alpha1-oleate containing system are coloured magenta, and the alpha1-Apo system coloured grey. (Pg. 36)

Figure 3.4. Per residue Secondary structure propensity calculations for Alpha1-oleate containing system and Alpha1 peptide in its apo form. The DSSP algorithm contains 8 definitions of representative secondary structure. To simplify our analysis, the 3-10 helix, α helix, and π helix structures were classified as "Helix"; the β-sheet, and residues in isolated β-bridge structures were classified as "Beta structures"; while the structures bend, coil, and turn were classified as "Others". A., B. Helical per residue propensity for Alpha1-oleate containing system and Alpha1-apo system, respectively. C., D. Beta structure per residue propensity for Alpha1-oleate containing system and Alpha1-apo system, respectively. E., F. "Others" secondary structure per residue propensity for Alpha1-oleate containing system and Alpha1-apo system, respectively. (Pg. 39)

Figure 3.5. Free energy surface as a function of the first two dihedral principal components and representative structures of identified local minima. A. Free energy surface as a function of the first two dihedral principal components for the Alpha1-oleate containing system. B. Free energy surface as a function of the first two principal components for the Alpha1-apo system. Minima A1, A2, A3, A4, A5 and A6 corresponds to local minima from the Alpha1-oleate containing system; while Minima B1, B2, B3, B4, B5 and B6 corresponds to local minima from the Alpha1-apo system. The Alpha1 peptide is coloured cyan, with the cyan sphere indicating the N terminus. Oleate molecules are shown as green sticks. (Pg. 41)

Figure 3.6. Representative conformations from geometric clustering of HAMLET Alpha1 peptide for the Oleate containing system (A) and Apo system (B). Alpha1 peptide shown as cyan cartoon. Oleate molecules shown as green sticks. The N-Terminal of the Alpha1 peptide is shown as a cyan sphere. (Pg. 44)

Figure 3.7. Contact probabilities between Alpha1 peptide residue sidechain protons and olefinic protons of any oleate molecule. A contact was defined if the minimum distance between side chain protons of residues from the Alpha1 peptide and olefinic protons was less than 0.45nm, and contact probability was calculated over the total number of simulation frames sampled in the last 900ns of H-REMD simulation. (Pg. 46)

Figure 3.8 Contact probabilities between heavy atoms of Alpha1 peptide residues and heavy atoms of individual oleate molecules in the folded state. Each oleate molecule is assigned a label oleate1-4. For contact probability calculation, a contact is defined if the minimum distance between the heavy atoms of oleate molecules and heavy atoms of each residue in the Alpha1 peptide is less than 0.45nm. Contact probabilities of more than 0.03 are shown. (Pg. 48)

Figure 4.1 Sequence alignment between human L-PGDS (PTGDS Human) (P41222) and Lipocalin 2 (NGAL human) (P80188). Sequence alignment performed on uniport. Asterisk (*) indicates identical residues, while colon (:) indicates residues with similar chemical properties. (Pg. 55)

Figure 4.2. Visualization of crystal structures of L-PGDS and Anticalin US7. A. Crystal structure of L-PGDS in its apo form (PDB ID 4IMN). B. Crystal structure of Anticalin US7 with Aβ (16-28) resolved (PDB ID 4MVI). C. Superimposition of Anticalin US7 structure with Aβ (16-28) resolved onto the structure of L-PGDS. L-PGDS is shown as green cartoon. Anticalin US7 is shown as yellow cartoon. Aβ peptide shown as purple cartoon, with N-terminus shown as a purple sphere. (Pg. 56)

Figure 4.3 Convergence of H-REMD simulation for L-PGDS- Aβ40 complex. A. Diffusiveness of H-REMD replicas were plotted for a representative low (effective temperature= 300K, black), medium (effective temperature= 414.57K, red) and high (effective temperature= 600K, blue) effective temperature replica. B. Minimum distance between heavy

atoms of L-PGDS and Aβ40. C.-E. Secondary structure propensities for Coil, Bend and Turn secondary structure, respectively as a function of simulation time. (Pg. 62)

Figure 4.4 Heat map of contact probabilities between residues of L-PGDS and Aβ40. Contact probabilities were calculated based on a minimum distance cutoff of 0.45nm between heavy atoms of L-PGDS and Aβ40 residues. Contact probabilities were calculated over the total number of simulation frames from 100ns-600ns of H-REMD simulation. The contact probability ranges from 0 (white) to 1 (dark blue) is shown. (Pg. 64)

Figure 4.5. Top 6 binding modes of L-PGDS towards Aβ40 and their percentages with respect to the total simulation frames population. (Pg. 66)

Figure 4.6 Representative binding pose of L-PGDS (Green) towards Aβ40 (Purple). A. Binding Mode 1, B. Binding Mode 2. C. Binding Mode 3. The N-Terminus of Aβ40 is shown as purple spheres. Residues identified to be involved in biomolecular recognition are shown as sticks. (Pg. 69)

Figure 4.7 RMSD of high contact probability residue pairs as a function of time for A. Binding Mode1, B. Binding Mode2, and C. Binding Mode3. Each of the three simulation repeats are coloured black, red and blue, respectively. (Pg. 71)

Figure 4.8. Contact map of average heavy atom distances between residues of L-PGDS and Aβ40. The average distances were calculated from the last 150ns of three simulation repeats for Binding Mode 1. The average distance was shown with a range of 0 (white) to beyond 1 nm (dark blue). Residues of L-PGDS which were identified by NMR to have undergone CSPs were highlighted in red, and Aβ40 residues which were identified by NMR to have undergone CSPs were highlighted in yellow. (Pg. 75)

Figure 4.9. Residues experimentally identified to have undergone chemical shift perturbation through NMR were highlighted red (L-PGDS) and yellow (Aβ40). L-PGDS is shown as green cartoon. The N-Terminal of L-PGDS is shown as green spheres. Aβ40 peptide shown as purple cartoon. The N-Terminal of Aβ40 is shown as purple sphere. (Pg. 77)

Figure 4.10. Calculated Solvent Assessible Surface Area (SASA) values per of L-PGDS in its apo form for residues that were experimentally identified to have perturbation in chemical shifts. SASA was calculated using gromacs tools over the average of three simulation repeats for the last 150ns. (Pg. 77)

Figure 4.11 Histogram of Rg distribution between L-PGDS apo (blue) and L-PGDS- Aβ40 complex (orange). Rg values were calculated from 3 200ns classical molecular dynamics simulations and using values from the last 150ns of each simulation repeat. The Rg values of

the L-PGDS- Aβ40 complex were calculated from the classical molecular dynamics simulations initiated from binding mode 1. (Pg. 79)

Figure 4.12. Superimposition of L-PGDS (Green Cartoon) -Aβ40 Binding Mode 1 model onto the crystal structure (PDB ID: 4MVI) of Anticalin US7 (Yellow Cartoon) in complex with Aβ (Purple Cartoon). For clarity, Aβ40 from Binding Mode 1 model is shown as cyan cartoon. N-Terminus of Aβ peptide shown as spheres (Pg. 80).

Figure 5.1. Sequence Alignment of the 5 AtPRFs isoforms. Alignment performed by clustal Omega. Identical residues are highlighted in red. (Pg. 84)

Figure 5.2 Crystal structures of AtPRFs. A. AtPRF3Δ37. B. AtPRF3Δ30SMplusL136Q. C. AtPRF2-PolyP complex. D. Superimposition of AtPRF2 in complex with PolyP (C) onto AtPRF3Δ30SMplusL136Q (B) reveals structure clash between N-Terminal Extension (NTE) of AtPRF3Δ30SMplusL136Q (B) (magenta cartoon) and PolyP binding site of AtPRF2. AtPRF3 shown as cyan cartoon, AtPRF3Δ30SMplusL136Q NTE shown as magenta cartoon, AtPRF2 shown as green cartoon, AtPRF2-PolyP shown as green sticks. N-terminal of PolyP shown as green spheres. (Pg. 87)

Figure 5.3. A. Free energy surface to characterize the dynamics of the N-Terminal Extension. It was plotted as a function of dihedral angle formed by C-Alpha carbons of residues 36 to 39 (Dihedral) against center of mass distance of the N-Terminal Extension and C-terminus helix (Distance). The free energy surface was generated through Hamiltonian Replica Exchange Molecular Dynamics simulation using the last 700ns of simulation data. Seven local minima of low relative free energies were identified and labelled 1-7. B. Representative structures of each local minima (cyan) was obtained through geometric clustering and superimposed onto AtPRF2-PolyP crystal structure (green). For clarity, only the PolyP segment of AtPRF2-PolyP crystal structure was shown (green sticks). The minima can be classified to be an “open” (yellow, minima6), “semi-open” (purple, minima 5 and 7), “closed” (red, minima 1-4) based on whether there is a clash of the NTE with the PolyP segment. Percentages refer to the ratio of the number of frames belonging to a particular minima to the total number of frames we sampled. (Pg. 97)

Figure 5.4. Convergence of H-REMD simulation by plotting various time blocks of the free energy surface, from A. 100ns to 550ns, B.100ns to 650ns, and C. 100ns to 800ns. (Pg. 99)

Figure 5.5. Binding pose of AtPRF3Δ37-PolyP. This was generated by superimposing AtPRF3Δ37 onto AtPRF2-PolyP complex. AtPRF3Δ37 shown as cyan cartoon. Aromatic residues in the profilin core shown as sticks and labelled. PolyP shown as green cartoon. Proline residues in close proximity of AtPRF3 aromatic residues are shown as green sticks, suggesting the identification of CH- π interactions. (Pg. 101)

Figure 5.6. Binding mode analysis of PRF3Δ22 variants in complex with PolyP. A. AtPRF3Δ22SM-PolyP cluster1. B. AtPRF3Δ22SM3RtoA-PolyP cluster 1. C. AtPRF3Δ22-PolyP cluster1. D. AtPRF3Δ22-PolyP cluster2. The AtPRF3 Profilin core shown as cyan cartoon. At PRF3 NTE shown as magenta cartoon. Residues of the AtPRF3 NTE identified to be involved in biomolecular interactions shown as magenta sticks. PolyP shown as green sticks. (Pg. 104)

Figure 5.7 PSIPRED server prediction of AtPRF3 secondary structure (Pg. 107)

Figure 5.8. Binding pose analysis AtPRF3Δ22SM-PolyP complex with NTE modelled in a partial helical conformation. A. AtPRF3Δ22SM-PolyP cluster1. B. AtPRF3Δ22SM-PolyP cluster2. C. AtPRF3Δ22SM-PolyP cluster3. (Pg. 108)

List of Tables

Table 3.1 Calculated Overall Secondary Structure propensity for Alpha1-oleate containing system and Alpha1-apo systems (Pg. 37)

Table 4.1 High contact probability residue pairs between L-PGDS and Aβ40 with their respective contact probabilities (Pg. 65)

Table 4.2 Three identified binding modes (BM) of L-PGDS- Aβ40 interaction (Pg. 67)

Table 4.3 Table of Contact Occupancies for identified biomolecular interactions in each binding mode of L-PGDS towards Aβ40 (Pg. 70)

Table 4.4 Binding Free Energy calculations for each Binding Mode (BM) of L-PGDS towards Aβ40. Energies in units of KJ/Mol (Pg. 72)

Table 5.1 Summary of various AtPRFs systems discussed in this chapter (Pg. 90)

Table 5.2 Calculated contact occupancies of identified biomolecular interactions in AtPRF3Δ22-NTE variants complexed with PolyP (Pg. 105)

Table 5.3. Binding Free energy Calculations using MM/PBSA for AtPRF3 variants in complex with PolyP. Units in KJ/mol (Pg. 106)

Table 5.4 Binding Free energy Calculations using MM/PBSA for PolyP bound AtPRF3Δ22SM with NTE partial helical conformation. Units in KJ/mol (Pg. 109)

List of Abbreviations

Alpha 1 peptide derived from the first Alpha helical domain of Human Alpha-lactalbumin (Alpha1)

Alzhiemer's disease (AD)

Amyloid beta peptide (Aβ40)

Arabidopsis thaliana (At)

Arabidopsis thaliana Formin Homology 1 Poly-Proline segment (PolyP)

Arabidopsis thaliana Profilin 3 (AtPRF3)

Arabidopsis thaliana Profilin 3 N-Terminal Extension (AtPRF3 NTE)

Binding Mode (BM)

Cerebro-spinal fluid (CSF)

Chemical Shift Perturbation (CSP)

Circular Dichroism (CD)

Classical Molecular Dynamics (cMD)

Define Secondary Structure of Proteins (DSSP)

Dihedral Principal Component Analysis (DPCA)

Formin Homology 1 (FH1)

General Amber Force Field(GAFF)

Gronigen Machine for Chemical Simulations (GROMACS)

Hamiltonian Replica Exchange Molecular Dynamics (H-REMD)

Heteronuclear Single Quantum Coherence (HSQC)

Human Alpha-lactalbumin Made Lethal to Tumor cells (HAMLET)

Intrinsically Disordered Proteins (IDPs)

Intrinsically Disordered Regions (IDRs)

Lipocalin Type-Prostaglandin D synthase (L-PGDS)

Molecular Dynamics (MD)

Molecular Mechanics (MM)

Molecular Mechanics energies combined with the Poisson-Boltzmann and Surface Area continuum solvation (MM-PBSA)

N-Termius Extension (NTE)

Nuclear Magnetic Resonance (NMR)

Nuclear Overhauser Effect (NOE)

Post-Translational Modifications (PTMs)

Principal Component Analysis (PCA)

Profilins (PRFs)

Quantum Mechanics (QM)

Radius of Gyration (Rg)

Replica Exchange Molecular Dynamics (REMD)

Replica Exchange with Solute Scaling (REST2)

Replica Exchange with Solute Tempering (REST)

Root Mean Square Deviation (RMSD)

Small Angle X-Ray Scattering in Solution (SAXS)

Solvent Accessible Surface Area (SASA)

Abstract

Proteins are an important class of biomolecules essential for life. They are involved in many biological functions, such as cell signalling, transport, defence mechanisms etc. Proteins must undergo two processes, namely folding and biomolecular recognition, in order to effectively perform their functions. Molecular Dynamics Simulations have been demonstrated as a successful tool used complementary to experimental methods in the understanding of biological processes. The projects in this thesis involve the application of Molecular Dynamics simulations in the study of various aspects of protein folding and biomolecular recognition.

The first project involves a study of biomolecular folding. The Human Alpha-lactalbumin Made Lethal to Tumor cells (HAMLET) complex was discovered as a tumoricidal complex of Human Alpha-lactalbumin and oleic acid cofactor in the ratio of 1:4-7. Experimental studies showed that the smallest subunit of Human Alpha-lactalbumin capable of forming tumoricidal complexes with oleic acid is its Alpha helical domains. In this project, we used Hamiltonian Replica Exchange Molecular Dynamics (H-REMD) simulations to sample the conformation space of the peptide derived from the first Alpha helical domain (Alpha1) of Human Alpha-lactalbumin in the presence of 4 oleic acid molecules. We also set up a parallel simulation system of the Alpha1 peptide alone in its apo form for comparison. In this study, we have unravelled conformational differences of the Alpha1 peptide in the presence and absence of the oleic acid cofactor.

In the second project, we studied the biomolecular recognition of the protein Lipocalin Type-Prostaglandin D synthase (L-PGDS) towards the amyloid beta peptide (Aβ40). L-PGDS has been identified to contain Amyloid beta chaperone activity, with the ability to bind to Aβ40 monomers and prevent their deposition and aggregation. L-PGDS was also found to contain

dis-aggregase activity towards pre-formed amyloid fibrils. Experimental evidence for the binding of L-PGDS to Aβ40 was detected in the form of residues that had significant chemical shift perturbations through Nuclear Magnetic Resonance (NMR) Spectroscopy. We aimed to construct a model detailing the interactions between L-PGDS and Aβ40. We adopted a biophysics-based approach by using H-REMD enhanced sampling and extracted high probability binding modes of the L-PGDS-Aβ40 complex through a clustering method based on high contact-probability residue pairs. Subsequent classical molecular dynamics simulations were performed to evaluate the stability of each binding mode and characterize various biomolecular interactions. This study provides the first model of L-PGDS in complex with Aβ40. We have identified and proposed several biomolecular interactions between L-PGDS and Aβ40, that can be further characterized experimentally. Such models of the interactions between L-PGDS and Aβ40 provides mechanistic insights into the biomolecular recognition functions of L-PGDS towards its binding partner.

In the third project, we studied the plasticity of the *Arabidopsis thaliana* Profilin 3 (AtPRF3) N-Terminus Extension (NTE) and its role in enhancing binding towards Formin poly-proline (PolyP). Profilins are proteins involved in actin cytoskeleton assembly. In addition to actin binding, the association of Profilins with Formin proteins plays regulatory roles in actin cytoskeleton polymerization. Of the five *Arabidopsis thaliana* (At) Profilin isoforms, AtPRF3 is found to be unique and expressed with an additional intrinsically disordered NTE that contribute to enhanced binding towards PolyP. Crystal structures of AtPRF3 in complex with PolyP was unable to be resolved experimentally. Moreover, the crystal structure of AtPRF3 in its apo form with residues K31 to N37 of the NTE resolved revealed a clash of the NTE with the PolyP binding site when superimposed onto the experimentally resolved crystal structure of AtPRF2 in complex with PolyP. We used H-REMD to sample the conformation space of the AtPRF3 NTE and demonstrated the possibility of the NTE to adopt “open” and “closed”

conformations to allow or occlude the binding of PolyP. Subsequently, we used classical molecular dynamics simulations and binding free energy calculations to understand the role of the NTE towards enhanced PolyP binding. We showed that the NTE can adopt multiple, non-specific, adaptive binding modes towards PolyP, consistent with experimental binding affinity measurements and the conformationally flexible nature of intrinsically disordered regions.

1. Introduction

1.1 Introduction to Proteins

Proteins are an important class of biomolecules essential for life. They constitute approximately 20% adult human body composition[1]. Proteins are involved in many critical biological processes, such as catalysis of biological reactions, transport, cell signalling, and cell defence mechanisms etc.[2]. The loss of protein function is implicated in many pathological conditions such as neurodegenerative diseases [3] and cancer [4]. Studies in the structure and function of proteins have thus involved many different disciplines of biology, and become important to our understanding of biological processes involved in maintaining life.

Proteins are composed of building blocks known as amino acids. Each amino acid contains an alpha carbon, linked to an amino NH2 group, a carboxyl COOH group, and a side chain (Fig. 1.1A). The side chain of an amino acid confers it it's specific chemical properties. The 20 naturally occurring amino acids may be classified according to their side chain properties as being non-polar, polar and charged. Each individual amino acid unit is joined by a peptide bond (Fig. 1.1B) during a condensation reaction, which results in the removal of water. The terminal of a protein may be defined based on the presence of the amino or carboxyl group at the free end, which are termed the Amino (N-) and Carboxyl (C-) terminus respectively (Fig. 1.1B)[2].

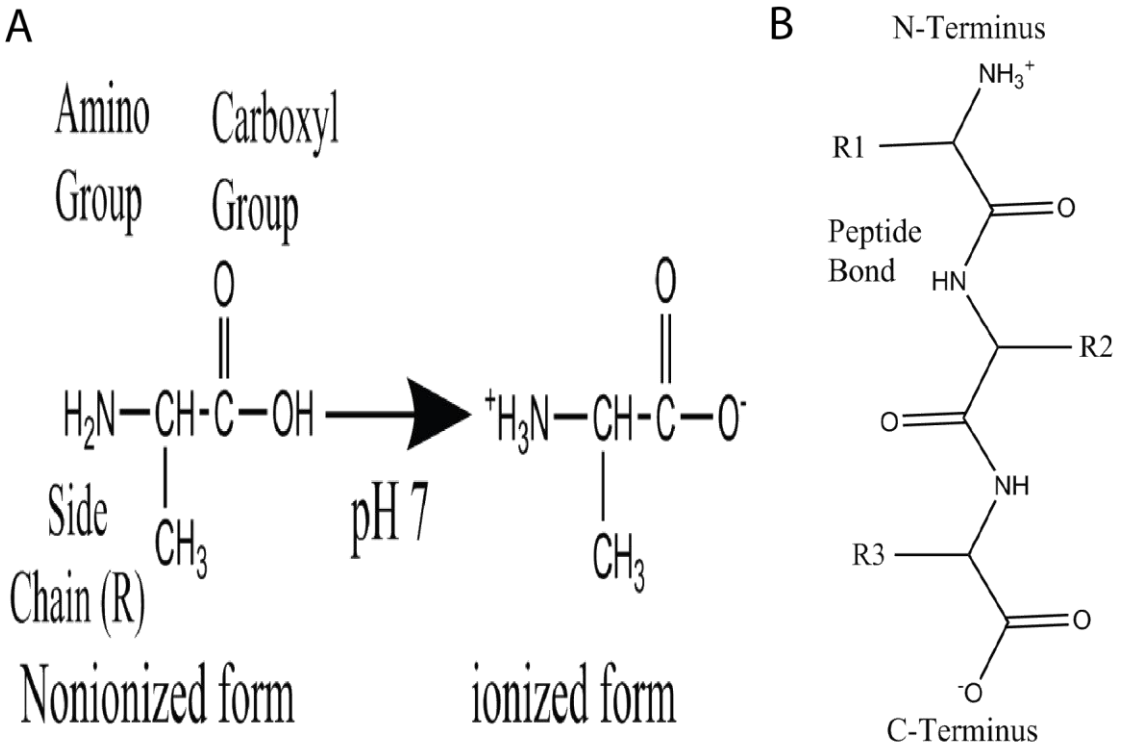

Figure 1.1. Schematic overview of an amino acid and the amino acid sequence of a short peptide. A. Schematic overview of an amino acid, showing the alpha carbon, amino and carboxyl groups, and the side chain. At pH7, the termini of amino acids are ionized and charged. B. Schematic overview of the amino acid sequence of a short peptide, showing an example consisting of 3 different amino acids (side chains labelled "R1", "R2" and "R3"). The peptide bond is labelled. The N- and C- terminals are indicated, respectively. Figure adapted from reference [2].

The structure of proteins may be classified hierarchically into 4 different levels, namely the primary, secondary, tertiary and quaternary structures[5]. The primary structure refers to the linear sequence of amino acids linked by peptide bonds in various different combinations. The secondary structure refers to regions of local order in the protein, and arises because of hydrogen bonding on the protein backbone. Two common types of secondary structures are the Alpha helix and Beta sheets[6]. Alpha helices are helical structures that form because of backbone hydrogen bonds from the i^{th} to the $i+4^{th}$ residue. Beta sheets form because of hydrogen bonding between the backbone of neighbouring Amino acid strands. The beta sheet is considered a parallel beta sheet if the amino acid sequence of the strand increases the same

orientation relative to the neighbouring strand. On the other hand, the beta sheet is considered to be anti-parallel if the amino acid sequence of the strand increases in an opposite orientation relative to the neighbouring strand. The tertiary structure of proteins refers to the folded three-dimensional conformation of a protein. This conformation is held by various biomolecular interactions, which may be covalent in nature, such as disulphide bonds, or weak non-covalent interactions. The quaternary structure of proteins involves the interaction of multiple subunits of protein chains associating with each other. A textbook example describing quaternary structure is haemoglobin[7], in which each molecule composes of multiple subunits of protein chains.

1.2 Introduction to Protein Folding

The process by which a protein acquires its tertiary, 3-dimensional structure is known as protein folding. Anfinsen, who won the 1972 Nobel prize in chemistry, conducted early studies on protein folding. In studies of ribonuclease A, beta-mercaptoethanol and urea was used to break the disulphide bonds and denature the tertiary structure of the protein, resulting in the loss of protein function. Upon using dialysis to remove the beta- mercaptoethanol and urea, the protein refolded and regained biological activity. The removal of beta-mercaptoethanol but not urea resulted in an inactive ribonuclease A, whereby its tertiary conformation was formed by random disulphide bond formation. When urea was removed and a trace amount of beta-mercaptoethanol was added, the inactive ribonuclease A refolded into an active conformation with biological activity[8]. Taken together, these experiments proposed several classical principles of protein folding. The amino acid sequence of a protein (Primary structure) determines the native state of a protein. The native state of a protein is considered a functional, unique and most stable conformation [9]. Also related to protein folding is Levinthal's paradox. The conformational degrees of freedom in a protein increases with the number of amino acids

in the sequence, and the time taken for proteins to fold to its native state would be very large if protein folding occurs via conformational sampling in a random and sequential manner [10]. To resolve this apparent paradox, conformational space of a protein may be described with free energy landscape theory [11]. Protein folding is proposed to be driven by the formation of local interactions that further drive protein folding process. This free energy landscape is likened to a funnel with a deep and narrow well, with the native state of a protein at the bottom with the lowest free energy. In the regions of high free energies, proteins can adopt a diverse range of conformations, and may undergo different pathways to reach the native state. Also, the rugged nature of the free energy landscape proposes that there might be intermediate conformational states that a protein may adopt during the folding process before reaching the native state [12]. The mis-folding of proteins into conformations other than the native state results in non-functional conformations. Misfolded proteins represent another low free energy state on the folding free energy landscape other than the native state in which proteins can fold into and remain stable. In fact, many neurodegenerative diseases, such as Alzhiemer's disease and Parkinson's disease can arise as a result of misfolded proteins, which deposit and aggregate into toxic oligomers [13].

Over the years, new discoveries expanded our understanding of biomolecular folding. One such discovery is that of Intrinsically Disordered Proteins (IDPs). IDPs are defined as proteins without the presence of a defined, ordered three-dimensional structures[14]. They can exist as either intrinsically disordered throughout their full length, or as segments of the globular protein being disordered, known as Intrinsically disordered regions (IDRs)[14]. IDPs and IDRs are found to be abundant in occurrence. It is predicted that approximately 25% of the eukaryote proteome is intrinsically disordered, while approximately 50% of eukaryote proteome contains IDRs[15]. Despite the absence of well-defined structure, it is shown that these proteins do

indeed play important roles in critical biological functions, such as in the regulation of cellular processes, cell signalling, and transcription, to name a few[16]. The discovery of IDPs and IDRs challenges the classical structure-function paradigm, in which the natively folded conformation is the functional conformation of a protein [9]. The presence of conformational plasticity and flexibility in IDPs/IDRs is hypothesized to confer these proteins evolutionary advantages that allows them to bind to multiple interaction partners [17]. IDPs/IDRs are characterized by the ability to adopt multiple conformations with similar energy levels that may rapidly interconvert. In the free energy landscape theory, IDPs and IDRs are characterized by the presence of wide and shallow energy minima [18].

Another discovery that adds new insights to our understanding of Protein folding is the discovery of the Human Alpha-lactalbumin Made Lethal to Tumor cells (HAMLET) complex. HAMLET is a tumoricidal complex of partially unfolded alpha-lactalbumin and oleic acid in the ratio of 1:4-7 [19]. Human Alpha-lactalbumin has a biological function of lactose synthesis in human milk production[20]. However, upon partial unfolding and in complex with oleic acid, it acquires a novel and distinct function of tumoricidal activity [19]. The discovery of HAMLET is proposed to represent a different category of protein folding, whereby proteins may undergo a conformational change to acquire a novel, beneficial function with the requirement of the presence of a co-factor[21].

1.3 Biomolecular recognition

For proteins to mediate their biological functions, they are required to interact with their binding partners. Such binding partners may include other biomolecules, or small organic molecules as cofactors. The process and mechanisms by which biomolecules interact with their binding partners is known as biomolecular recognition. Biomolecular recognition brings about

biomolecular function. For example, the binding of proteins to their binding partners may be part of a signalling cascade that result in further binding events, eventually leading to a biological response. Another example of biomolecular recognition is enzyme catalysis. Enzymes that catalyses biological reactions contain a well-defined binding pocket, which serves to position the substrate and co-factors for the reaction in close proximity[2].

Biomolecular recognition may be explained by three mechanisms, namely lock-and-key, induced fit and conformational selection mechanisms. The lock and key hypothesis proposed by Fischer in 1894[22] suggests that shape complementarity between two biomolecules result in them fitting well together. This theory assumes that biomolecules are static in their conformations. The induced fit model suggests that upon binding of the complex, the biomolecules involved would alter their conformations to result in a suitable fit [23]. The conformational selection model suggests that binding between biomolecules selects for a specific conformation in the populations of conformations, resulting in a population shift[24]. Both induced fit and conformational selection theories are not mutually exclusive. There are many experiments and simulations in support of both theories [25, 26]. This is because it may be possible for both induced fit and conformational selection to play roles in the biomolecular recognition process. For example, in the case of the Lysine-, Arginine-, Ornithine-binding (LAO) protein, the conformational selection was proposed to be responsible for the formation of an encounter complex as the apo form of the protein was shown to be able to conformationally access this state. Subsequently, the induced fit mechanism results in a conformational change to the ligand-bound state by interactions of the protein with the ligand[27].

Biomolecular recognition is generally driven by the hydrophobic effect, which is dominantly entropic in nature. Briefly, water molecules form a “clathrate cage” structure around hydrophobic groups, and the ordering of water results in an entropy penalty that eventually affects the free energy to make a biological process unfavourable[28]. Hydrophobic groups tend to aggregate because the aggregation of hydrophobic groups minimizes the disruption of bulk water hydrogen bonding network, and minimizes the entropy penalty due to ordering of water molecules around a solute [29]. It is noted that this phenomenon is non-specific in nature. The entropic effect generally drives biomolecular recognition in solvent during the initial encounter process, while the specificity aspect of biomolecular recognition relates to the binding enthalpy[30]. The binding enthalpy component relates to the formation of favourable biomolecular interactions between a biomolecule and its binding partners. This is commonly exploited and applied in rational drug design to optimize favourable biomolecular interactions between a biomolecule and a drug candidate. Through the analysis of Isothermal Titration Calorimetry data, it is shown that second and third generation of HIV Protease inhibitors bind with a strong enthalpic contribution. This was attributed to the large number of backbone-specific biomolecular interactions between the inhibitor and the target[31]. Specificity arises due to the presence of favourable biomolecular interactions between binding partners. Such interactions comprise weak, non-covalent interactions, for example, electrostatic interactions, hydrogen bonding, aromatic ring interactions such as aromatic ring stacking [32]. The identification of biomolecular interactions in this thesis is covered in detail in (Section 2.6 Identification of Biomolecular Interactions).

1.4 Molecular dynamics simulations for studies of biomolecular folding and recognition

Molecular Dynamics (MD) simulations have become an important and powerful tool in the study of many biological processes[33]. This technique allows us to study biomolecular dynamics at an atomistic scale of resolution. Briefly, the MD algorithm solves Newton's equations of motion for a particular biological system[34]. In doing so, the ensembles of conformations that can be adopted by the system at specific conditions, such as temperature and pressure, are generated[35]. MD has thus become a useful tool in conformational space sampling[33]. Molecular dynamics of biomolecules was pioneered in the 1970s by McCammon with the simulation of bovine pancreatic trypsin inhibitor (BPTI)[36]. Advances in the development of high-performance computing and developments in computational algorithms have enabled larger system sizes as well as longer simulation timescales length to be simulated. For example, D.E Shaw research performed a 1 ms simulation of BPTI[37] and human ubiquitin[38]. This was done through the use of a specialized built supercomputer, Anton[39] for the purpose of MD simulations. However, because of the lack of access to such specialized hardware, it is more common for MD simulations to be performed at shorter timescales. This can be done using enhanced sampling methods, for example, Replica Exchange Molecular Dynamics involving multiple simulation replicas, or performing multiple short simulation trajectories to develop Markov State Models that describes a system's dynamics[40]. A recent example involves several thousand short simulations of 100-200ns of bacterial ribonuclease barnase with its inhibitor barstar, resulting in 2ms aggregate length of simulation sampling[41].

MD simulations provide several advantages which experimental methods cannot provide. MD simulation provides information about the interactions and movements of the individual atoms

in the system. Often, experimental measurements of macroscopic properties are ensemble averaged values. Also, most experimental methods do not have as high a resolution to allow an adequate understanding of biological processes at the atomistic level[42].

Also, certain structural biology methods may be challenging to perform for proteins which are conformationally flexible, such as IDPs or IDRs. For example, X-Ray Crystallography, a common structural biology technique to study the three-dimensional conformations of biomolecules, requires the packing of biomolecules in a crystal lattice. The conformationally flexible nature of IDPs and IDRs makes the process of obtaining well diffracting crystals and clear electron density a challenging process[43]. Also, the crystallization conditions, such as buffer, salt, pH etc. may also result in artificially packed conformations of a biomolecule[43]. Often, for X-Ray crystallography studies, the flexible regions and domains of the biomolecule may be truncated, or mutations of residues performed in order to stabilize regions/domains to be crystallized. Such structural features components may be critical to biological process of interest[43]. Another common structural biology technique to study of biomolecular interactions is Nuclear Magnetic Resonance Spectroscopy (NMR). The common solution NMR techniques are limited by protein size[44]. Also, conformation exchange in biomolecules results in broadening of NMR peaks and loss of resonance signals, making the peak assignment in the NMR spectra and characterization of ligand binding challenging[44]. In the absence of experimental structures, molecular modelling of biomolecular folding and recognition to understand the biomolecular interactions between a biomolecule and it's binding partners can performed using in-silico MD simulations[42]. This constitutes a major theme in this thesis.

Moreover, biomolecules are not static entities. Conformational changes and dynamics are important for their function. For instance, by superimposing crystal structures of proteins in

their apo forms and ligand bound forms, different types of structural changes, ranging from small active site rearrangements to large domain movements may be observed [45]. Molecular dynamics simulations allow the generation of an ensemble of conformations with control of temperature and pressure[33], being advantageous compared to structural biology methods that only provide information about a single snapshot. Study of individual structures alone is inadequate to explain biological phenomenon. The conformational flexibility of biomolecular systems and their conformational changes are important in understanding biomolecular processes can be captured in MD simulations[33].

MD simulations are commonly used in complement together with experimental results. Experiments can guide MD simulations by proposing research questions and serve to validate the findings of MD simulations. On the other hand, simulations can be used to confirm experimental hypotheses, as well as propose mechanistic explanations for observation of biological phenomenon. Thus, MD simulations allow us to understand and interpret experimental results that cannot be explained by experiment alone. Also, MD simulations allows new predictions to be made that can be verified experimentally. This approach is commonly used in drug discovery or mutagenesis studies[33].

1.5 Aims and scope of the thesis

This thesis aims to study various aspects of Biomolecular folding and recognition through the use of Molecular Dynamics Simulations.

Chapter 2 provides a review on the methods applied in this thesis. The methods covered are classical molecular dynamics simulations, enhanced sampling methods, and binding free

energy calculations. An overview of the identification of biomolecular interactions is also covered.

Chapters 3 to 5 covers the research that was undertaken for three different projects during the Ph.D study, in which the themes focuses on various aspects of Protein folding and biomolecular recognition. Each chapter consists of an introduction, which provides an overview of biological system and biological research question. Following which, the specific methodology and system simulation parameters are detailed. Results, discussions and conclusions are also presented in each individual chapter for each specific project.

In chapter 3, we aim to study the biomolecular folding of the peptide (Alpha1) derived from the first Alpha helical domain of Human Alpha-lactalbumin, in complex with oleic acid cofactor. This is part of our study in understanding the Human Alpha-lactalbumin Made Lethal to Tumor cells (HAMLET) complex, a tumoricidal complex of Human Alpha-lactalbumin with oleic acid co-factor in the ratio of 1:4-7. By comparison to a simulation system of the Alpha1 peptide alone, we demonstrated conformational differences between the Alpha1 peptide in the presence and the absence of the oleic acid cofactor. Based on our observed conformational differences between the Alpha1 peptide in the presence and absence of oleic acid, we may propose a hypothesis for the structural basis of the tumoricidal activity of the Alpha1-oleic acid complex and suggest a role for the oleic acid co-factor.

In chapter 4, we study the biomolecular recognition between the protein Lipocalin Type-Prostaglandin D synthase (L-PGDS) and the amyloid beta peptide (Aβ40). The protein L-PGDS was discovered to contain beta chaperone activity. The direct binding of L-PGDS to Aβ40 monomers functions to prevent the deposition and aggregation of Aβ monomers, hence

this binding constitutes a form of biomolecular recognition. In this study, we used enhanced sampling and extracted high probability conformations of L-PGDS- Aβ40 complex. We then show the agreement of our simulation identified binding modes with experimental data. This study provides the first model of the L-PGDS- Aβ40 complex. Such a model provides mechanistic value to understanding the biomolecular recognition of L-PGDS towards it's binding partners.

In chapter 5, we studied the plasticity of the *Arabidopsis thaliana* Profilin 3 (AtPRF3) N-Terminus Extension (NTE) and its role in biomolecular recognition towards Formin poly-proline (PolyP). The association of Profilins with Formin proteins at their proline rich regions plays important roles in regulation of actin cytoskeleton polymerization. AtPRF3 is a unique profilin isoform expressed with an intrinsically disordered NTE. Using H-REMD simulations, we demonstrated the conformational plasticity of the AtPRF3 NTE, and show that it has the potential to occlude or expose the PolyP binding site to allow or prevent PolyP binding. We also used classical MD simulations and binding free energy calculations to study the biomolecular recognition of AtPRF3 NTE towards Formin PolyP, and show that the NTE is able to adopt multiple, non-specific binding modes in its interaction with PolyP.

Finally, chapter 6 provides the conclusion and future outlook to works presented in this thesis.

2. Method

2.1 Introduction to Molecular Modelling

Molecular modelling involves the use of mathematical approximations to construct models of the system of interest. Two common approaches are used to model biomolecular systems, namely the Quantum mechanics (QM) approach and molecular mechanics (MM) approach.

The QM method involves solving the Schrodinger equation for the system, providing information about the wavefunction and energy. This itself is an approximation, since an analytical solution to the Schrodinger equation is available for single electron systems. The QM method provides relatively accurate energy calculations compared to the molecular mechanics method. They have been shown to be useful where quantum effects are important [46]. However, there are some disadvantages of QM methods. Firstly, QM calculations are limited by a small system size (approximately 200 atoms) because of the computational costs involved. Also, calculations are performed in gas phase or implicit solvent model [47]. While MD based on QM is quite popular, for example through the use of CPMD[48] or CP2K[49] software, the treatment of atoms using QM increases the computational expense substantially and impose limits on system size.

For biomolecular systems such as proteins, where the system size is an important consideration, it is considered impractical to use QM approaches for modelling. On the other hand, molecular mechanics (MM) methods involves treating atoms as beads held by springs and point charges. In this modelling, electrons are not considered. As such, quantum effects such as covalent bond formation or transition metal chelation cannot be considered. The energy of the system is defined by an empirical function. This allows the energies of a large biological system with

explicit solvent to be approximated rapidly. This reasonable approximation allows the study of biological processes at a longer timescale[35].

2.2 Force Field

The potential energy function may be expressed as a set of empirical parameters, is known as the force field[50]. This is used to construct a model and description of the interactions between atoms in a biomolecular system. These parameters that represent the geometries and interaction between particles are derived from theoretical quantum mechanical calculations or fitting to experimental values. The choice of force field parameters is important for the accuracy of molecular modelling. Examples of commonly used force fields are the CHARMM[51], AMBER[52], GROMOS[53], and OPLS[54] force fields. Force fields differ in the ways they are parameterized and their functional forms. New force fields are constantly being developed or re-parameterized and from new experimental values to provide more accurate parameter sets for molecular modelling[55]. An example is the recent CHARMM36m[56] force field, which is parameterized for folded and Intrinsically Disordered Proteins.

The Potential energy can be classified into bonded and non-bonded terms. Bonded interactions include bond stretching, bond bending, and torsion about a dihedral angle. Non-bonded interactions include van der Waals dispersion and repulsion, and coulombic electrostatic interactions between atoms. The potential energy of the system V(R) can be expressed as a sum of bonded and non-bonded terms (equation 2.1):

$$V(R) = E_{bonds} + E_{angles} + E_{torsions} + E_{vdw} + E_{columb} \quad (2.1)$$

$$V(R) = \sum_{bonds} k_b \left(l - l_0\right)^2 + \sum_{angles} k_a \left(\theta - \theta_0\right)^2 + \sum_{torsions} \frac{1}{2} V_n (1 + \cos(n\omega - \gamma)) + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \epsilon_{ij} \left[\left(\frac{r_{0ij}}{r_{ij}}\right)^{12} - 2\left(\frac{r_{0ij}}{r_{ij}}\right)^{6} \right] + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} k \frac{q_i q_j}{r_{ij}} \quad (2.2)$$

The first 3 terms in the potential energy function (equation 2.2) describes the bonded interactions in a system. The first term E_{bonds} refers to the energy due to the variation of bond lengths from an equilibrium value, and a harmonic potential is used to describe it. The term l_0 refers to the equilibrium value of bond length. The term $k_b$refers to the harmonic force constant. The second term E_{angles} refers to the energy resulting from the deviation of bond angles from an equilibrium value. The term θ_0 is the angle equilibrium value. k_a is the angle bending force constant. The third term $E_{torsions}$ describes the energy resulting from the alteration of dihedral angles. 4 consecutively bonded atoms rotate about a plane defined by the central bond. This is described by a Fourier series, with term V_n describing the torsional barrier height and the term n relates to the periodicity. The 4th and 5th terms of equation 2.2 describe non-bonded interactions in a system. The 4th term E_{vdw} describes the van der Waals interactions between atom pairs. This is modelled by a 12-6 Lennard Jones potential. The term r_{0ij} refers to the equilibrium distance and the term ϵ refers to the well depth. The fifth term E_{columb} represents the electrostatic interaction between atom pairs. The terms q_i and q_j represents point charges of two atoms i and j, while the term r_{ij} refers to the distance between two atoms i and j. The term k is the electrostatic constant[50].

2.3 Classical Molecular Dynamics simulations

Molecular dynamics (MD) simulations involves solving Newton's equations of motion for a particular system in order to produce a time dependent simulation trajectory of the particles' movement. This can be used to provide atomistic level detail of the motions of particles[34]. Briefly, an initial biomolecule structure provides coordinates which describes atoms initial positions. The forces are computed for each atom based on the potential energy function defined by the force fields, and atoms are then assigned initial velocities from a Maxwell-Boltzmann distribution. Newton's equation of motion is solved and each atom's position and

velocity are updated at each timestep. These steps are repeated in a loop continuously for the required simulation duration.

The potential V is described as a function of the position of particles,

$$V = V(\overrightarrow{r_1},\ \overrightarrow{r_2},\ \ldots.\overrightarrow{r_n}) \tag{2.3}$$

Newton's equation of motion describes the force acting on a particle i as follows,

$$F_i = -\nabla_i V(R) = m_i a \tag{2.4}$$

$$a = \frac{dv}{dt} = \frac{d^2 R_i(t)}{dt^2} \tag{2.5}$$

F_i is the force acting on atom i as a function of atomic coordinates, $R_i$and m_i refers to the coordinates and mass of atom i, respectively. R refers to the set of positioning vectors of atoms. The term a refers to the acceleration, which is related to the derivative of velocity v and position, that results in a change in the positions and velocities over time (equation 2.5). A small time step Δt in the order of the fastest motions of the system is used. Newton's equations of motion can be integrated through the use of the leap-frog algorithm[57], which is implemented in the Gromacs (Gronigen Machine for Chemical Simulations) [58] Molecular Dynamics software package,

$$R(t + \Delta t) = R(t) + v(t + \tfrac{1}{2}\Delta t)\Delta t \tag{2.6}$$

$$v\left(t + \tfrac{1}{2}\Delta t\right) = v\left(t - \tfrac{1}{2}\Delta t\right) + \frac{F_i(t)}{m_i}\Delta t \tag{2.7}$$

The position of the subsequent time step $R(t + \Delta t)$ (equation 2.6) is calculated from the current position $R(t)$ and velocities at the next half time step $v\left(t + \frac{1}{2}\Delta t\right)$ (equation 2.7). This can be calculated from the current acceleration $\frac{F_i(t)}{m_i}$ and the velocities at the previous half time step

$v\left(t-\frac{1}{2}\Delta t\right)$. The force $F_i(t)$ and position $R(t)$ can be calculated with velocities at the previous half time step $v\left(t-\frac{1}{2}\Delta t\right)$. The current velocities may be calculated from,

$$V(t) = \frac{1}{2}\left(v\left(t+\frac{1}{2}\Delta t\right) + v(t-\frac{1}{2}\Delta t)\right) \quad (2.8)$$

Thus, the future positions and velocities are determined from the current positions and velocities[57].

2.4 Hamiltonian Replica Exchange Molecular Dynamics

One limitation of classical molecular dynamics (cMD) simulations is the sampling problem. Many biological processes occur at a timescale not accessible to cMD simulations. For example, local side chain rotation occurs in the nanosecond timescale, while large domain changes, ligand binding and folding events can occur in the microsecond, millisecond or second timescale. Also, high energy barriers between minima may result in local minima traps for cMD simulations. This may lead to an inadequate sampling of the conformation space. Hence there is a need for enhanced sampling methods that have been developed to overcome the limitations of cMD simulations[59, 60].

The Hamiltonian Replica Exchange Molecular Dynamics (H-REMD) method is an example of an enhanced sampling method that can be used to overcome sampling limitations. In this method, several parallel replicas of the system are simulated with differing potential energies. At a certain time interval, replicas would attempt to exchange conformations based on a specified exchange criterion. The Hamiltonian of a system may be decomposed to allow scaling of specific components in order to reduce energy barriers and allow greater energy overlaps that would in turn enhance the probability of exchange.

There have been many diverse implementations of H-REMD. The general principle involves the modification of the Hamiltonian to reduce potential energy barriers [61]. An example by Kannan and Zacharias involved the use of a peptide backbone biasing potential to enhance peptide backbone transitions between replicas by lowering backbone dihedral transition barriers[62]. Hritz and Oostenbrink used soft-core potentials to weaken strong repulsion forces between atoms close in space to reduce energy barriers[63]. The Replica Exchange with Solute Scaling (REST2)[64] method and its first iteration Replica Exchange with Solute Tempering (REST)[65] are examples of scaling potential energies through modification of force field parameters. A comprehensive reference on the different H-REMD methods may be found in [66]. The REST2 method was demonstrated to be greatly improve sampling efficiency in proteins which undergo large scale conformation changes[64]. It is also implemented for ease of use in the GROMACS molecular dynamics software with PLUMED plugin[66].

In the Replica Exchange with Solute Scaling (REST2) method, the system's total interaction energy can be written as follows,

$$E = E_{PP} + E_{PW} + E_{WW} \tag{2.9}$$

The term E_{PP} refers to the intra-molecular energy of the protein. The term E_{PW} refers to the interaction energy between the protein and water. The term E_{WW} refers to the self-interaction energy between water molecules. In this implementation of H-REMD[64], the potential energy of each replica is scaled by a scaling factor as follows for a particular replica m,

$$E_m^{REST2} = \frac{\beta_m}{\beta_0} E_{PP}(X) + \sqrt{\frac{\beta_m}{\beta_0}} E_{PW}(X) + E_{WW}(X) \tag{2.10}$$

The term $\beta_m = 1/K_b T_m$, where T_m is temperature at replica m and T_0 refers to the biologically relevant ensemble temperature, in this case 300k. It is important to note here that simulations for all replicas are carried out at 300k. The "effective temperature" term for REST2 is used to

generate the scaling factors for the potential energy of replica. A scaling factor $\frac{\beta_m}{\beta_0}$, a number smaller than 1 reduces the energy barriers between replicas and aids in enhancing sampling. Analysis would only be performed on the replica which the potential energies are unscaled. The square root in scaling factor for Epw was proposed by Terakawa et. al.[67] and implemented in REST2[64]. This was proposed by the authors to optimize the perturbed Hamiltonian to allow efficient conformational exchange[67].

The exchange between the m th and n th replica may be described by the following expressions[65]:

$$(X_m, E_m(X_m), T_m) \rightarrow (X_n, E_m(X_n), T_m)$$
$$(X_n, E_n(X_n), T_n) \rightarrow (X_m, E_n(X_m), T_n) \quad (2.11)$$

In this expression, $X_m, E_m(X_m), T_m$ are the configuration, energy and temperature of the mth replica just before an exchange is attempted. Corresponding terms are used to describe properties of the nth replica before the exchange.

The equilibrium probability for the mth state is defined by following equation,

$$P_m = \frac{1}{Z_m} \exp(-\beta_m(E_m(X_m)) \quad (2.12)$$

In equation 2.12, $\beta_m = 1/K_b T_m$ and Z_m is the configurational partition function.

The transition probability of the forward exchange can be specified by the expression T($i \rightarrow$ F) and reverse exchange T(F$\rightarrow$ i). By imposing the detailed balance condition, the relationship between transition probabilities of the forward and reverse exchange is described by the following relationship,

$$P_m(X_m)P_n(X_n)T(i \rightarrow f) = P_n(X_m)P_m(X_n)T(f \rightarrow i) \quad (2.13)$$

The ratio of transition probabilities can be described as follows:

$$\frac{T(i \rightarrow f)}{T(f \rightarrow i)} = \exp(-\Delta_{nm}) \quad (2.14)$$

Where $\Delta_{nm} = -\beta_m[E_m(X_n) - E_m(X_m)] - \beta_n[E_n(X_m) - E_n(X_n)]$ (2.15)

By substituting equation 2.10 into equation 2.15,

$$\Delta_{mn}\ (REST2) = (\beta_m - \beta_n)\left[(E_{PP}(X_n) - E_{PP}(X_m)) + \frac{\sqrt{\beta_0}}{\sqrt{\beta_m}+\sqrt{\beta_n}}(E_{PW}(X_n) - E_{PW}(X_m))\right]$$

(2.16)

Δ_{mn} is the acceptance ratio. The acceptance ratio for exchange between replicas m and n is related to the energy differences between conformations at different replicas as well as Hamiltonian scaling. The advantage of using H-REMD in this implementation is that the term E_{WW} for water-water self interaction energy is not considered in the acceptance ratio formula. It is shown that this approach is able to reduce the number of replicas as compared to Temperature-Replica Exchange Molecular Dynamics yet increase the sampling efficiency[64].

By applying the metropolis criterion, the acceptance probability may be described as:

$$T(i \rightarrow f) = \begin{cases} 1\ if\ \Delta_{nm} \leq 0 \\ \exp(-\Delta_{nm})\ if\ \Delta_{nm} > 0 \end{cases} \quad (2.17)$$

This means that if $\Delta_{nm} \leq 0$, the acceptance probability is 1. If $\Delta_{nm} > 0$, a random number from 0 to 1 is drawn and compared to $\exp(-\Delta_{nm})$. If $\exp(-\Delta_{nm})$ is greater than the random number, the exchange is accepted.

The choice of scaling factors for each replica is important because they eventually affect the exchange acceptance probability between replicas. Low exchange acceptance probabilities between replicas would result in inefficient sampling. In this thesis, the scaling factors were chosen through the use of a geometric progression and an adequate number of replicas. Short simulations prior to production simulations were performed to optimize the scaling factor range and the number of replicas to result in a reasonable average exchange acceptance probability of at least approximately 0.25.

Unlike other enhanced sampling methods such as the metadynamics method[68] which involves adding a bias potential along certain collective variables of a system, the H-REMD method does not require the defining of reaction collective variables *a-priori*[59]. Identification of collective variables, which describe the reaction coordinates of a system is a challenging process[69]. However, the use of H-REMD results in discontinuous trajectories because of conformational exchange between replicas. Also, this method is computationally expensive because of the need for multiple replicas with differing potential energies to enhance the sampling of the replica with unscaled potential energies[60].

2.5 Binding Free Energy Calculations

The MM-PBSA (molecular mechanics energies combined with the Poisson-Boltzmann and surface area continuum solvation) method is to estimate the binding free energy between a protein and its binding partner. A comprehensive review of the method can be found in Genheden and Ryde, 2015[70].

The binding free energy is defined as $E_{Binding} = E_{vdw} + E_{elec} + E_{Polar} + E_{Apolar} - T\Delta S$, whereby E_{vdw} refers to van der Waal's interaction energy, E_{elec} refers to electrostatic interaction energy, E_{Polar} refers to polar solvation energy approximated by the Poisson

Boltzmann equation, E_{Apolar} refers to the Apolar solvation energy approximated using a Solvent Accessible Surface Area (SASA) model, T refers to temperature in K, while ΔS refers to entropy[70]. The entropy can be estimated with the quasi-harmonic approach[71]. To calculate the absolute binding free energy, the entropy term is an important consideration[70]. The MM-PBSA method shows good correlation between calculated binding free energies values with experimental measurements[72]. However, calculation of entropy is computationally expensive and also filled with convergence issues. Hence, a common approach is to perform the MM-PBSA calculations without the entropy term[73] to calculate binding free energies. Such a calculation allows a comparison between systems without the need for absolute free energy calculations. In our systems, our aim of binding free calculations is to perform a relative comparison between similar systems. The systems compared are almost identical in nature with exception of a few residue mutations. Moreover, our binding free energy calculations are supported by experimental microscale thermophoresis assay measurements, which provide a reference to the relative order of magnitude difference in energy values due to residue mutations.

2.6 Identification of biomolecular interactions

Biomolecular interactions refer to the weak, non-covalent attractive forces between biomolecules. Though weak by themselves, the force of attraction between biomolecules can be substantial when combined. In biomolecular recognition, specificity is important to ensure the cognate binding partners for a biomolecular process bind to each other. The array of specific biomolecular interactions between biomolecules contribute to specificity of biomolecular recognition process[2].

Since biomolecular interactions arises due to properties of electrons, Quantum Mechanical calculations should be used to analyse and characterise biomolecular interactions [74]. However, because of the limitations of system size and computational cost requirement for Quantum Mechanics methods (elaborated in section 2.1 Introduction to Molecular Modelling), we use classical approximations to identify biomolecular interactions. The typical distances for various biomolecular interactions have been reviewed in literature[32, 75, 76]. A recent study by Piovesan et al involved extensive analysis of X-ray crystallography and NMR spectroscopy structures to understand the distance distributions of various biomolecular interactions[76]. In this thesis, we use a series of distance measurements to identify and characterize biomolecular interactions. Firstly, the non-bonded terms in the force field, the electrostatic and van der Waal's interactions are approximated as a function of distance (Section 2.2 Force Field). Moreover, the approach of using distance to identify biomolecular interactions is common in the field of structure based drug discovery, in which biomolecular interactions are identified from the three-dimensional structure of a biomolecule for the purpose of rational drug design to enhance interactions between a receptor and ligand[31]. It is noted that such an identification of biomolecular interactions is a simplification because effects of entropy is not considered. However, the analysis of entropy is computationally expensive and difficult to converge[71], and entropic effects are not as intuitively identified as enthalpic effects. Also, the presence of solvation reduces the overall magnitude of the biomolecular interactions compared to that in-vacuo. Solvation results in a screening effect which reduces the range of electrostatic interactions[77].

We seek to identify biomolecular interactions from our simulation generated models. Such information may be further utilized by experimental collaborators to drive further experiments, for example, mutagenesis studies. In addition, the nature of MD sampling[34] allows us to be

able to calculate the occupancies of identified biomolecular interaction as a function of the total number of simulation frames we sampled.

In the force field, biomolecular interactions between biomolecules are grouped as non-bonded interactions, and may be broadly classified as electrostatic and van der Waal's interactions (Section 2.2 Force Field). Electrostatic attraction is an attractive force that occur between atoms of opposite charges[78], while Van der Waal's interactions refer to the sum of the weak intermolecular forces that include dipole or induced dipole mediated interactions[79]. Examples of biomolecular interactions which we identify in this thesis for specificity are electrostatic interactions, hydrogen bonding, aromatic ring (π- π) stacking, cation-π stacking and CH- π interactions.

Electrostatic interactions refer to the attractive force between atoms of the opposite charges. In proteins, these interactions occur between amino acids with charged side chains. The charged amino acids at physiological pH are the Arginine, Lysine, Aspartate and Glutamate amino acids. It is suggested that electrostatic interactions may contribute to initial recognition between a biomolecule and it's binding partners due to its long range nature[75]. The strength of electrostatic interactions is related to the distance between two atoms (Section 2.2 force field). In this thesis, we approximate the identification of electrostatic interactions by a 0.5nm distance between oppositely charged atoms of charged Amino acid residues. The interaction energy of an electrostatic interaction is approximately 5-10Kcal/mol[75].

A special type of electrostatic interaction is the Hydrogen Bond. It involves an electronegative atom as a hydrogen bond donor, and a hydrogen bond acceptor, which is an electronegative atom with a lone pair of electrons. The hydrogen bond donor, which is electronegative and

covalently bonded to a proton, results in the proton having a partial positive charge character. In proteins, hydrogen bonds may occur between peptide backbones or amino acids with polar side chains. Hydrogen bonds are directional in nature. In this thesis, we used Gromacs[58] tool gmx hbond definition, which is based on a donor-acceptor atom distance of 0.35nm and Hydrogen-Donor-Acceptor cutoff angle of 30 degrees to define the presence of a hydrogen bond. The magnitude of a hydrogen bond is approximately 3-5Kcal/mol[75].

Aromatic ring (π-π) stacking arises because of the delocalization of electrons in π orbitals of amino acid residues with aromatic side chains (phenylalanine, tryptophan, tyrosine and histidine). Different forms of orientation of aromatic side chains, such as "Edge-to-Face" stacking and "Face-to-Face" stacking may be observed in aromatic ring stacking interactions[32]. For simplicity, we identify aromatic ring stacking interactions using a centre of mass distance between side chains of aromatic residues of 0.65nm[76]. Aromatic ring stacking interactions do not have explicit energy terms in the force field. However, they may be implicitly taken into account during force field parameterization process and may considered as van der waals interactions. As a model example, the dimerization energy of benzene is approximately -1.6 to -2.4Kcal/mol [80].

Cation- π interactions arises because of interaction between positively charged amino acids of a protein and amino acids with aromatic side chains (Phenylalanine, tryptophan, tyrosine and histidine residues). In this thesis, we approximate identification of Cation- π interactions by a distance of 0.5nm between the positively charged atom of positively charged amino acids and the centre of mass of aromatic residues side chains. There is no explicit terms in the force field for Cation- π interactions. This form of interaction may be considered a form of electrostatic interaction[81]. The magnitude of Cation- π interaction is approximately 2.6Kcal/mol[82].

CH–π Interactions arises between delocalized π electron systems in aromatic side chains of amino acids and Aliphatic bonds. They arise because of electron delocalization and dispersion effects. Multiple of such interactions is suggested to form a network to stabilize the interactions between a biomolecule and it's binding partners [83]. In this thesis, we approximate the identification of CH–π interactions with a distance of 0.55nm from Aliphatic carbon of residues with hydrophobic side chains to the centre of mass of aromatic residues side chains. There is no explicit terms in the force field for CH- π interactions. As a model system, the calculated benzene-methane interaction energy by MP2 QM calculation is approximately -1.45Kcal/mol [84].

3. In-silico modeling of the Human Alpha-lactalbumin Made Lethal to Tumor cells (HAMLET) Alpha1 complex

3.1 Introduction

3.1.1 HAMLET (Human Alpha-lactalbumin Made Lethal to Tumor cells) is a tumoricidal complex of Alpha-lactalbumin and oleic acid

Human Alpha-lactalbumin Made Lethal to Tumor cells (HAMLET) was reported in 1995 by a group of researchers from Lund University, Sweden. HAMLET was discovered during the use of a human lung cancer cell line during studies of how fractions of human milk affected bacterial cell adhesion. Upon treatment with human milk, cell death was observed on the cancer cells in an apoptosis like mechanism, but healthy differentiated cells were unaffected. A fraction of human milk contained tumoricidal activity, affecting a wide variety of tumor cell lines, while not affecting healthy, differentiated cells[85]. Subsequent research identified components of the HAMLET complex to be unfolded Human alpha-lactalbumin and oleic acid in a ratio of 1: 4-7 [19] (Fig. 3.1).

Human alpha-lactalbumin is a 123 amino acid residue protein with a biological function involved in the synthesis of lactose, and is the most abundant protein in human breast milk[20]. The crystal structure of Human Alpha-lactalbumin at 1.7Angstroms resolution was reported in 1991 by Acharya et al. In the structure, the protein contains four alpha helices and an anti-parallel beta sheet in between. Four disulphide bridges and a divalent calcium ion is found to stabilize this globular and compact conformation of Human Alpha-lactalbumin [86]. Upon conditions such as EDTA treatment or low pH that results in the loss of Ca2+, a molten globule state of Human Alpha-lactalbumin is formed, characterized by tertiary structure loss, maintained secondary structure, and increased hydrophobic domain exposure[87]. The

partially unfolded Alpha-lactalbumin binds with a lipid cofactor identified to be oleic acid, forming a complex that acquires tumoricidal activity[88]. The partial unfolding of Human Alpha-lactalbumin and undergoing of a conformation change is essential for the formation of the tumoricidal complex with oleic acid. The native, folded conformation of human alpha-lactalbumin is found to be incapable of forming tumoricidal complexes with oleic acid[89]. It has also been demonstrated that the tumoricidal effects of HAMLET are not achieved by Human Alpha-lactalbumin unfolding only, and it was necessary to introduce the oleic acid as a cofactor. Unsaturated 18 carbons fatty acids with a single cis double bond (Oleic acid), but not other fatty acid variants were able to form the tumoricidal HAMLET complex [90]. At the concentration where the HAMLET complex is tumoricidal, oleic acid alone is not tumoricidal[90]. Consistent with the requirement for partial unfolding of human alpha-lactalbumin, it was also found a reduced human alpha-lactalbumin mutant with all cysteines mutated to alanines could form a cytotoxic complex in the presence of the oleic acid lipid cofactor [91].

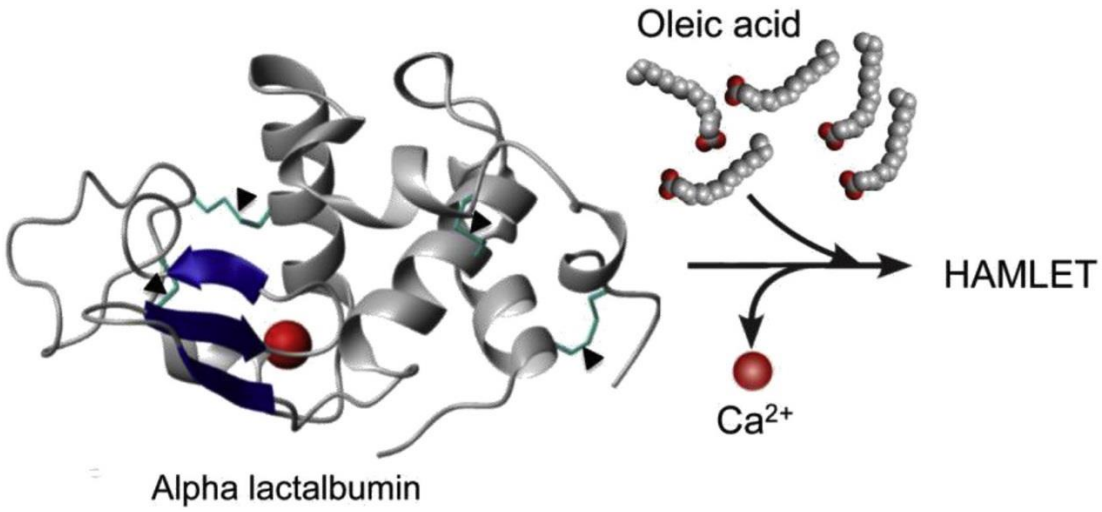

Figure 3.1 Introduction to the formation of Human Alpha-lactalbumin Made Lethal to Tumor cells (HAMLET) complex. Human Alpha-lactalbumin binds with oleic acid in the ratio of 1:4-7 and partially unfolds with the removal of Calcium to form the HAMLET complex. Figure adapted from reference [19].

3.1.2 Biological Effects of HAMLET

The Biological effects of HAMLET can be classified as a broad-spectrum apoptosis like response towards tumor cells without the affecting of healthy, differentiated cells[92]. HAMLET interacts with and alters cell membranes[89] in a receptor independent manner[93], triggering ion-fluxes that result in the subsequent activation of cell death signalling pathways[94]. Upon entry into tumor cells, the HAMLET complex induces apoptotic like morphological changes, including loss of cytoplasm and nuclear condensation[95]. It interacts with targets such as kinases and GTPases [96]. HAMLET was also found to co-localize with and inhibit proteasome activity[97]. In the mitochondria, cytochrome C release follows mitochondrial membrane depolarization and swelling [98]. HAMLET has demonstrated activity in more than 40 different cancer cell lines[99]. For in-vivo studies, HAMLET has demonstrated activity in Glioblastoma [100], Palpilloma[101], and bladder cancer[102] models. The interested reader is invited to pursue a recent review detailing the biological effects of HAMLET[19].

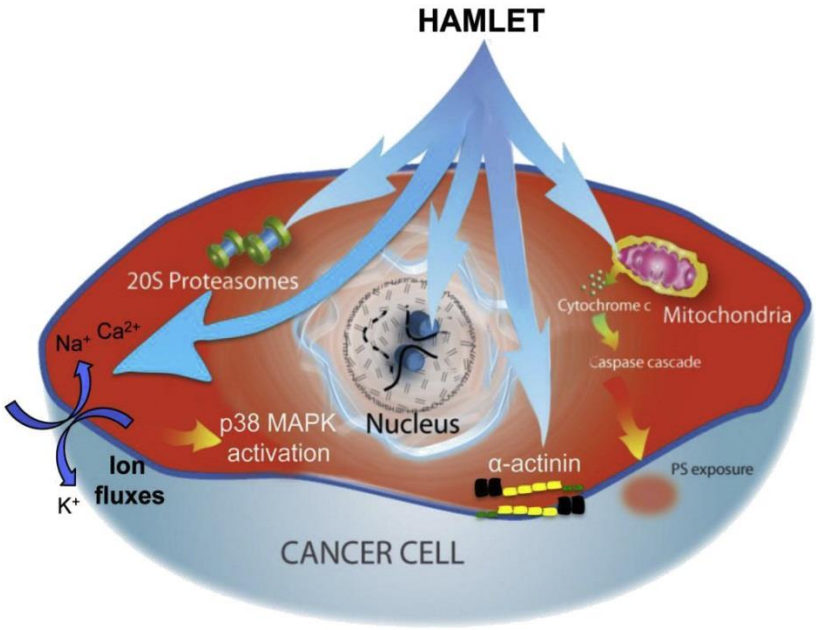

Figure 3.2 Figure summarizing the biological effects of Human Alpha-lactalbumin Made Lethal to Tumor cells (HAMLET). The Biological effects of HAMLET may be described as a broad-spectrum apoptosis like response towards tumor cells without the affecting of healthy, differentiated cells Figure adapted from reference [19].

3.1.3 Structural studies of the HAMLET complex

In 2012, Small Angle X-ray scattering in solution (SAXS) was used to solve the low resolution structure of HAMLET [103]. The SAXS data suggests no aggregation of the human alpha-lactalbumin and that Human alpha-lactalbumin monomer interacts with oleic acid. The space filling model showed that HAMLET comprises of a large globular core with an extended tail, demonstrating that Human Alpha-lactalbumin undergoes a conformation change to form the HAMLET complex with oleic acid. By superimposing the crystal structure of Human alpha-lactalbumin with the SAXS space filing model of HAMLET, it was proposed that the C-terminal residues from L105 to L123 of Human alpha-lactalbumin could adopt an extended conformation[103].

In the same study, the researchers attempted to identify the smallest possible region of Human Alpha-lactalbumin that could form a tumoricidal complex with oleic acid. This was done through generating peptides comprising the different domains of Human Alpha-lactalbumin-namely the N-terminal alpha helical domain (Alpha1); the C-terminal alpha helical domain (Alpha2) and beta domain in between the N- and C- terminal alpha helices (Beta1). Cysteine residues were mutated to Alanine residues to facilitate the partial unfolding of these peptides, consistent with a study that these mutations on Human Alpha-lactalbumin could result in a tumoricidal complex formed with oleic acid [91]. These peptides were complexed with oleic acid and tested for tumoricidal activity. It was found that the Alpha1-oleic acid complex and Alpha2-oleic complex were able to induce tumor cell morphology changes resulting in rounding up and detachment of cells. They were also able to cause Tumor cell death in A549 lung carcinoma cells and Jurkat leukemia cells. However, the Beta1-oleic acid complex did not contain tumoricidal activity. Performing the same assays with shorter 15 amino acid peptides corresponding to the sequence of Human alpha-lactalbumin did not result in the formation of

tumoricidal complexes with oleic acid. The smallest region of Human alpha-lactalbumin that could form tumoricidal complexes with oleic acid were the Alpha1 and Alpha2 alpha helical domains[103].

3.1.4 Research Aim and direction

Despite multiple functional studies performed characterizing the activity of HAMLET, little high-resolution conformational information is available regarding Human Alpha-lactalbumin and oleic acid in the HAMLET complex. There is a lack of understanding of the structural basis for the tumoricidal activity of the HAMLET complex. Moreover, the conformation differences between Human Alpha-lactalbumin in the presence and absence of oleic acid are still not clearly understood. We aim to use molecular modelling to investigate these aspects through molecular dynamics simulations.

Our simulation studies are performed using the peptide derived from the first alpha helical domain of human alpha-lactalbumin, in which the cysteines have been mutated to Alanines (Alpha1). This is because folding of a 123 amino acid residue full length human alpha-lactalbumin is computationally prohibitive given the limitations of computational power. Although the Alpha1 peptide is derived from the first alpha helical domain of human alpha-lactalbumin, the simulation of Alpha1 is not intended to be a replacement for a simulation of the full-length HAMLET complex. Experiments showing that the smallest segment of alpha-lactalbumin capable of forming a cytotoxic complex with oleic acid was found to be an Alpha domain, supporting the folding simulation study of the Alpha1 peptide with oleic acid. Moreover, the Alpha1 peptide in complex with oleic acid has been extensively characterized separately from the full-length HAMLET complex in terms of its tumoricidal activity [103].

In this research, we attempted to elucidate the conformational differences between the oleic acid bound form of the Alpha1 peptide and it's apo counterpart. We used Hamiltonian Replica Exchange Molecular Dynamics in the Replica Exchange with Solute scaling (REST2) variant[64], an enhanced sampling method to sample the conformational space of the Alpha1 peptide in the presence of 4 oleic acid molecules. For comparison purposes, we correspondingly set up an apo system of the Alpha1 peptide alone without the oleic acid. Through these simulations, we have gained insights into the structural differences between the lipid cofactor bound forms and the apo forms of the Alpha1 peptide. A key major difference identified was the Alpha1 peptide in the presence of oleic acid adopting conformations with well-defined alpha helical structures and the oleic acid molecules interacting in a hydrophobic core. In contrast, the Alpha1 peptide in the absence of oleic acid adopts a diverse variety of conformations with varying secondary structure.

3.2 Method

3.2.1 Initial model construction

The initial conformation of the Alpha1 peptide is taken from the first 39 residues of crystal structure of Human Alpha-lactalbumin at 1.15Angstroms resolution (PDB ID 1B9O) reported in 1999 by Harata et al [104]. All cysteines were converted to alanines, consistent with findings that a reduced Human Alpha-lactalbumin variant in which all cysteines mutated to alanines could form a tumoricidal complex in the presence of the lipid cofactor [91]. For consistency with experimental conditions, the Alpha1 N-Terminal was capped with COCH3 group. The Alpha1 peptide is placed center in a cubic box with 4 oleic acid molecules placed randomly around the Alpha1 peptide to obtain a peptide: lipid ratio of 1:4. A corresponding system with the Alpha1 peptide alone is also set up. Peptides were solvated in a cubic box using TIP3P [105] water model with a minimum distance of 1.2nm from solute and the box edge. All systems were neutralized and Na+ and Cl- ions were added to a concentration of 0.15 M. For

the peptide amino acid residues, Amber99SB-ILDN [106] force field was used. Oleic acid coordinates was built using Discovery Studio 4.1[107]. The oleic acid was modelled in a deprotonated oleate form. Topologies for the oleic acid were built using the General Amber Force Field (GAFF) [108]. Geometry optimization for the oleate was performed using Gaussian09[109] at the level of HF-6-31G*, and the partial charges were determined by the RESP[110] method implemented in the antechamber tool of AmberTools16. The LINCS[111] algorithm was used to constrain all bonds containing hydrogen. Energy minimization was performed using the steepest descent algorithm for 1000 steps to remove any initial bad contacts. Long-range electrostatics were treated with the particle mesh ewald algorithm[112]. A cutoff of 1.2 nm was implemented for electrostatics, and van der Waal's interactions were truncated at 1.2 nm. Both the Alpha1 peptide-oleate containing system and the system with the Alpha1 peptide alone were initially heated at 500 K for 40 ns to eliminate initial structure bias as well as render the peptides in a partially unfolded state. Temperature coupling of the system was performed using a velocity rescaling[113] thermostat.

3.2.2 Hamiltonian replica exchange molecular dynamics

The Gromacs[58] 5.1.2 molecular dynamics package with the Plumed[114] 2.3 plugin for Hamiltonian Replica Exchange Molecular Dynamics was used to perform the simulations, using the Replica Exchange with Solute Scaling (REST2) [64] variant. All atoms of the Alpha1-peptide residues, along with the oleate molecules of the oleate-containing system, were selected for Hamiltonian scaling. Twenty replicas were used for each system, and scaling factors were generated for an effective temperature range of 300 K to 800 K. Temperatures for scaling were selected based on a geometric progression. The temperatures were 300, 315.893, 332.629, 350.251, 368.807, 388.346, 408.919, 430.583, 453.395, 477.415, 502.707, 529.34, 557.384, 586.913, 618.006, 650.747, 685.223, 721.525, 759.75 and 800 K. For each system,

each replica was simulated for 1μs, resulting in an effective simulation of 20 μs. Exchanges were attempted every 2 ps, and the result was an average acceptance probability of approximately 25%. Simulation frames were saved every 2 ps.

3.2.3 Analysis of simulation data

The ensemble for each system with the canonical unscaled potential energy (replica with effective temperature= 300K) was used for analysis. To take into account the effects of initial equilibration, data analysis was performed on the last 900ns for each system with the exception of simulation convergence analysis.

Simulation data was analyzed using the in-built gromacs[58] tools. Hydrogen bond number was calculated using the gmx H-bond tool. Peptide radius of gyration was calculated using the gmx gyrate tool. Number of peptide-oleate heavy atom contacts was calculated using the gmx mindist tool. Dihedral Principal Component Analysis[115] was performed using the tools gmx angles, gmx covar, and gmx anaeig in order to prepare and diagonalize the covariance matrix, as well as perform analysis of eigenvectors and eigenvalues. For geometric clustering, the gmx cluster tool was used. In order to calculate secondary structure propensities, we used the Define Secondary Structure of Proteins (DSSP) algorithm, a tool which determines protein secondary structure based on hydrogen bond pattern analysis[116]. 8 definitions of representative secondary structure are defined by the algorithm, namely the 3-10 helix, α helix, π helix, hydrogen bonded turn, β-sheet, residue in isolated β-bridge, bend, and coil. For the purpose of our analysis, the 3-10 helix, α helix, and π helix structures were classified as helix, the β-sheet, and residue in isolated β-bridge structures were classified as beta structures, while the remaining structures were classified as others. Contact probability was calculated using the gmx mindist tool in the gromacs package. A contact is defined if the measured distance

between two groups is less than 0.45nm, and contact probability was calculated by expressing number of frames containing a particular contact over the total number of frames sampled in the last 900ns.

3.3 Results

To understand the conformational differences between the Alpha1-peptide in the presence and absence of oleate, we performed Hamiltonian Replica exchange molecular dynamics for 2 systems, one in the presence of oleate and the absence of oleate.

3.3.1 Assessment of simulation convergence

In order to assess simulation convergence, we monitored several system properties as a function of simulation time. First, we monitored the diffusiveness of representative H-REMD replicas as a function of simulation time, which showed adequate mixing of simulation replicas, for both the Alpha1-oleate containing (Fig. 3.3A) and Alpha1-apo (Fig. 3.3E) systems. We also plotted the i to i+4 hydrogen bond number of the Alpha1 peptide in the presence and absence of oleate molecules (Fig. 3.3.B, 3.3F), the Alpha1 peptide radius of gyration in the presence and absence of oleate molecules (Fig. 3.3C, 3.3G), as well as the peptide-oleate number of heavy atom contacts for the Alpha1 peptide-oleate containing system (Fig 3.3D). These properties show consistent fluctuation and do not change drastically with the increase in simulation time, thus demonstrating the reasonable convergence of our simulations.

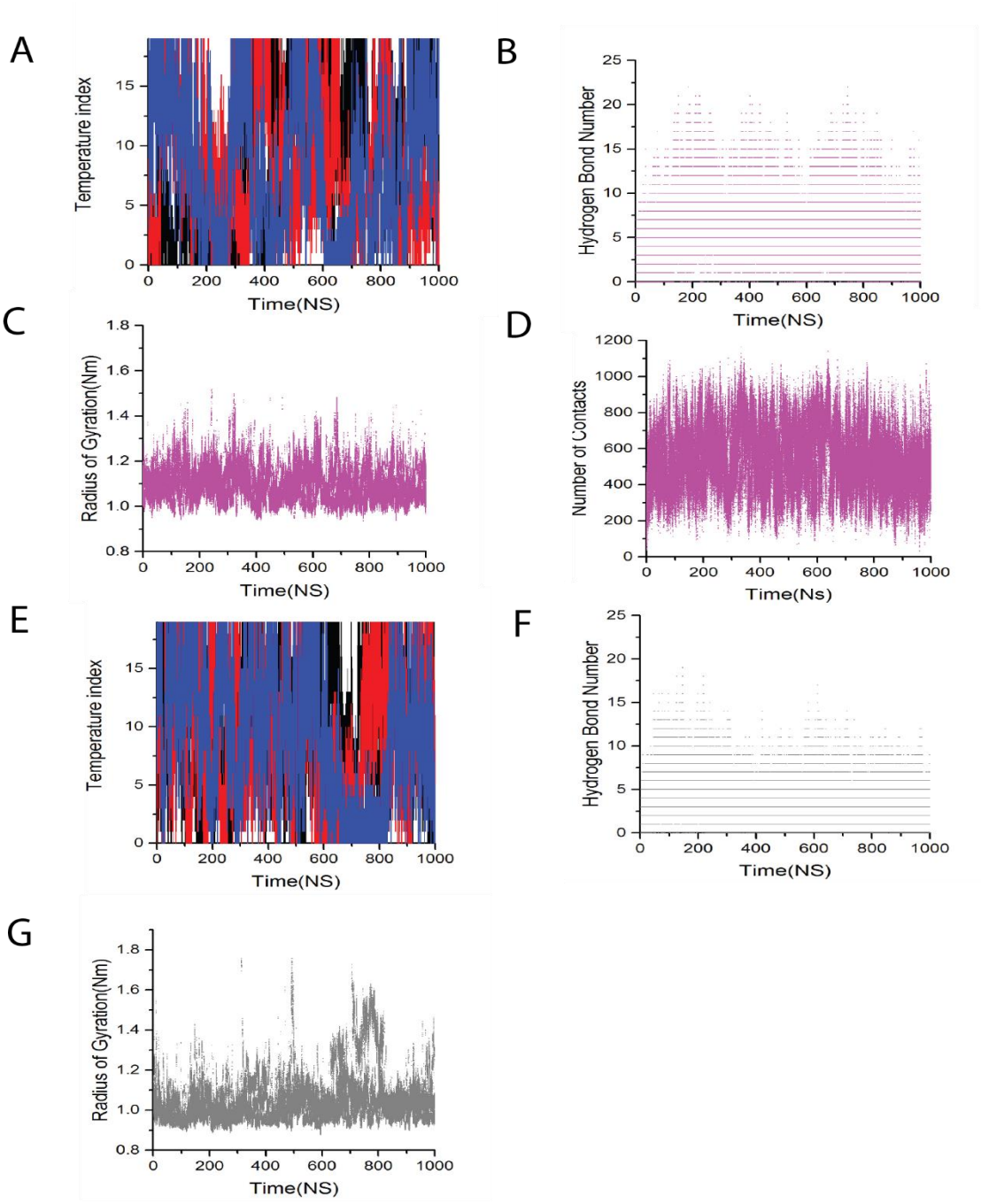

Figure 3.3. Evaluation of Hamiltonian Replica Exchange Molecular Dynamics simulation convergence. Diffusiveness of representative H-REMD replicas from a representative low (effective temperature= 300K, Black), medium (effective temperature= 477.4K, Red) and high (effective temperature= 800K, Blue) effective temperature replica for Alpha1-oleate containing system (A) and Alpha1-apo system (E). Simulation properties as a function of time were plotted. i to i+4 Hydrogen bond number for Alpha1-oleate containing system (B) and Alpha1-apo system (F). Alpha1 Peptide Radius of Gyration for the Alpha1-oleate containing system (C) and Alpha1-apo system (G). Number of heavy atom contacts between the Alpha1 peptide and oleate molecules (D) in Alpha1-oleate containing system. For clarity, properties relating to the Alpha1-oleate containing system are coloured magenta, and the alpha1-Apo system coloured grey.

3.3.2 Secondary structure analysis reveals higher percentage of helical conformations for the Alpha1-oleate containing system

To investigate the secondary structure differences between the Alpha1-oleate containing system and the Alpha1-apo system, we calculated the overall secondary structure propensity (Table 3.1) as well as the per residue secondary structure propensity (Fig. 3.4) of the Alpha1 peptide in both systems. The DSSP algorithm[116] contains 8 definitions of representative secondary structure, namely the 3-10 helix, α helix, π helix, hydrogen bonded turn, β-sheet, residue in isolated β-bridge, bend, and coil. To simplify our analysis, we classified the 3-10 helix, α helix, and π helix structures as "helix" structures, the β-sheet, and residues in isolated β-bridge structures as "beta structures", and the remaining secondary structure types structures as "others". The overall secondary structure propensity calculations for the Alpha1 peptide show a higher helical secondary structure propensity of 0.27 in the presence of oleate compared to a helical structure propensity of 0.18 for the Alpha1 peptide alone in the absence of oleate (Table 3.1). This finding agrees with Circular Dichroism spectroscopy measurements from our experimental collaborators, which suggests that the Alpha1-oleate containing system contained a relatively higher helical secondary structure propensity compared to the Alpha1 peptide in its apo form (Experiments performed by experimental collaborators from Lund University, Sweden).

Table 3.1 Calculated Overall Secondary Structure propensity for Alpha1-oleate containing system and Alpha1-apo systems

Secondary Structure	Alpha1- 4oleate	Alpha1- apo
Helix	0.27	0.18
Beta Structures	0.08	0.1
Others	0.65	0.72

To identify which residues were involved in helix formation, we calculated the per residue secondary structure propensity of the Alpha1 peptide for both the Alpha1-oleate containing system and the Alpha1-apo system. The per residue secondary structure plot of helical structures for the Alpha1-oleate containing system (Fig. 3.4A) shows 2 continuous segments of residues with per residue helical propensity greater than 0.2, with one segment from residues 5 to 12, and a second segment from residues 24 to 33. By contrast, the Alpha1 peptide in its apo form (Fig. 3.4B) showed shorter segments of residues with per residue helical structure propensity greater than 0.2, with one segment comprising residue 6 to 11, and another segment comprising residues from residues 25 to 30. Moreover, a closer examination of the per residue helical propensity of individual residues in the segments revealed a relatively higher per residue helical propensity in the Alpha1-oleate containing system compared to the Alpha1-apo system (Fig. 3.4A, Fig.3.4B). For instance, in the Alpha1-oleate containing system, segments of residues 8 to 10 and residues 25 to 28 of the Alpha1 peptide had a high per residue helical propensity of more than 0.6 (Fig. 3.4A). On the other hand, none of the per residue helical propensity exceeded 0.6 for the Alpha1 peptide in its apo form (Fig 3.4B). Low per residue beta structures propensities were observed for both the Alpha1 peptide in the presence (Fig. 3.4C) and absence (Fig. 3.4D) of Oleate. In addition to the N and C termini of the Alpha1 peptide, both systems show a region of high per residue propensity for disordered secondary structure at the region separating the 2 defined helical segments (Fig. 3.4E and 3.4F, classified as “others”). This could be explained by the presence of Glycine residues 17, 19, and 20 in the Alpha1 peptide sequence, which are known to be helix breakers or disruptors of helical conformations[117, 118].

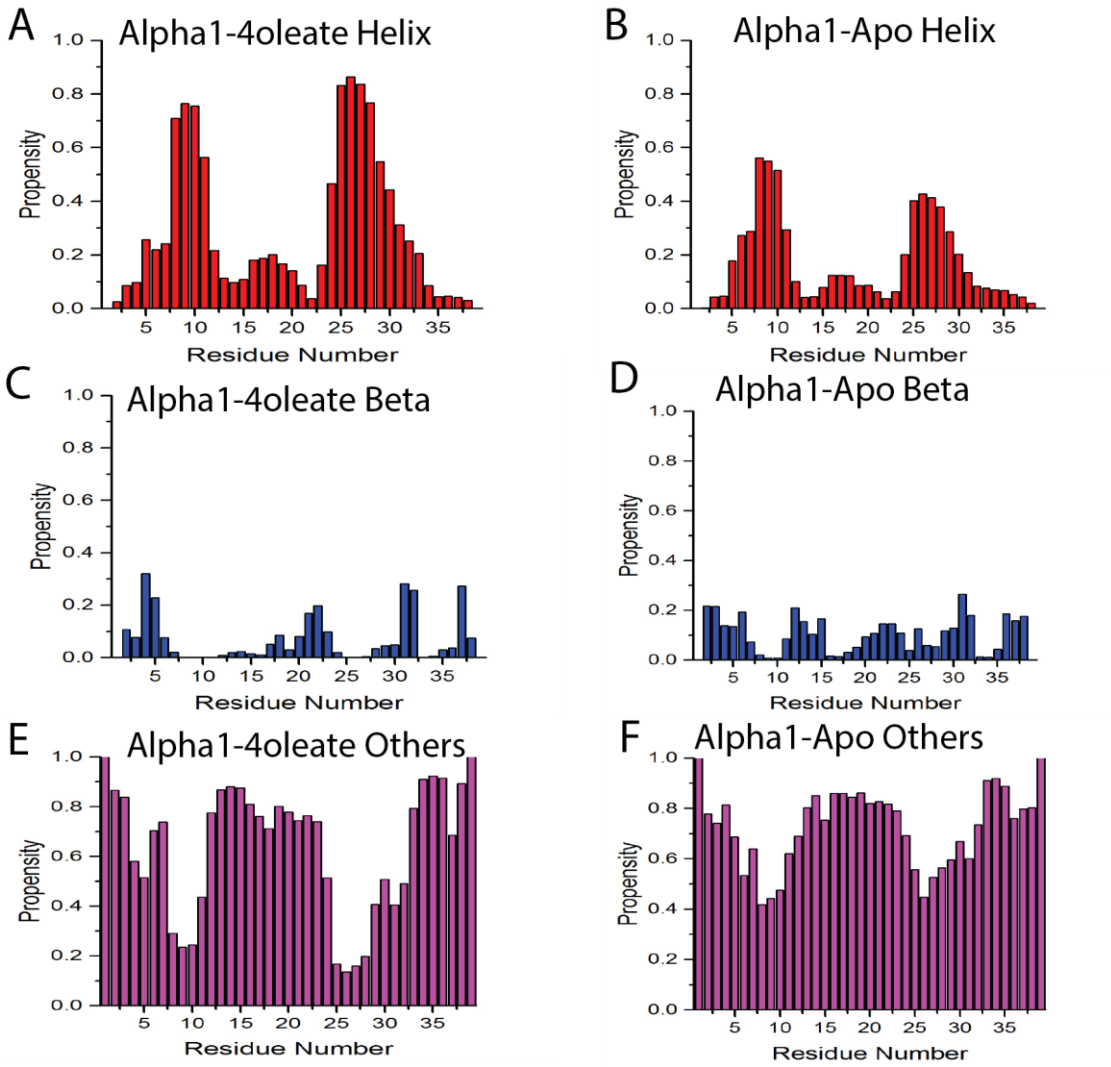

Figure 3.4. Per residue Secondary structure propensity calculations for Alpha1-oleate containing system and Alpha1 peptide in its apo form. The DSSP algorithm contains 8 definitions of representative secondary structure. To simplify our analysis, the 3-10 helix, α helix, and π helix structures were classified as "Helix"; the β-sheet, and residues in isolated β-bridge structures were classified as "Beta structures"; while the structures bend, coil, and turn were classified as "Others". A., B. Helical per residue propensity for Alpha1-oleate containing system and Alpha1-apo system, respectively. C., D. Beta structure per residue propensity for Alpha1-oleate containing system and Alpha1-apo system, respectively. E., F. "Others" secondary structure per residue propensity for Alpha1-oleate containing system and Alpha1-apo system, respectively.

3.3.3 Free energy landscape as a function of Principal components of dihedral angles reveal conformational differences between both systems

Having understood the secondary structure differences between the Alpha1-oleate containing system and the Alpha1 peptide in its Apo form, we sought to understand the conformational differences between both systems. Principal Component Analysis (PCA) is a method of dimension reduction in order to identify important degrees of freedom in a system. Dihedral Principal Component Analysis (DPCA) is a method developed to use the backbone dihedral angles as internal coordinates for a representation of the dynamics of a peptide[115]. We used DPCA to understand the major conformational differences between the Alpha1 peptide in the presence and absence of oleate. The dihedral angles of the Alpha1 peptide for both the oleate containing and apo systems were combined to prepare the covariance matrix in order to identify the axes of greatest variance, so that a meaningful comparison can be made. The free energy surface as a function of the first two Dihedral Principal Components was constructed through projection of dihedral angles of each system onto the first and second principal components. The plot is then divided into bins, and the relative free energies of each bin is calculated with the formula $F_i = -RT \ln(P_i/P_0)$, where the term R is the gas constant, T is temperature (300 K), P_i refers to the population in each bin, and P_0 is the population of the most populated bin. Geometric clustering was performed by the Gromos algorithm in the Gromacs[58] package to identify representative structures of each local minimum. The free energy surface, together with the representative structures of the identified local minima for the Alpha1-oleate containing system and the Alpha1-apo system are shown in Figure 3.5. From the free energy surface, it is apparent that the Alpha1-oleate containing system and Alpha1-apo system are located in different regions in subspace, with some common regions of overlap. This suggests that structures of the Alpha1 peptide in the presence and absence of oleate adopt diverse and different conformational ensembles (Fig. 3.5A, B).

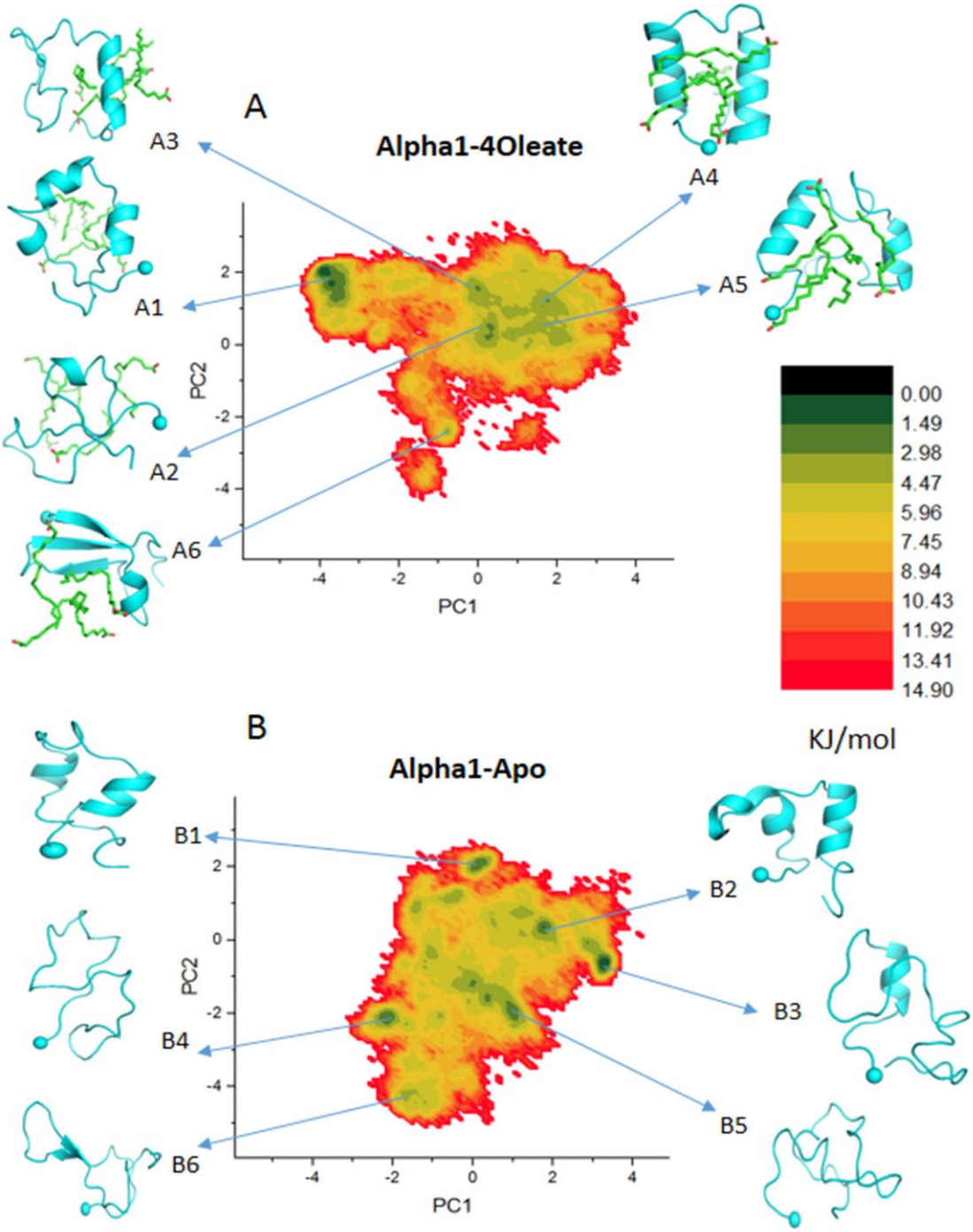

Figure 3.5. Free energy surface as a function of the first two dihedral principal components and representative structures of identified local minima. A. Free energy surface as a function of the first two dihedral principal components for the Alpha1-oleate containing system. B. Free energy surface as a function of the first two dihedral principal components for the Alpha1-apo system. Minima A1, A2, A3, A4, A5 and A6 corresponds to local minima from the Alpha1-oleate containing system; while Minima B1, B2, B3, B4, B5 and B6 corresponds to local minima from the Alpha1-apo system. The Alpha1 peptide is coloured cyan, with the cyan sphere indicating the N terminus. Oleate molecules are shown as green sticks.

The Free Energy Surface of the Alpha1-oleate containing system and the Alpha1-apo system is characterized by multiple minima separated by low free energy barriers (Fig. 3.5A, Fig. 3.5B). In the Alpha1-oleate containing system (Fig. 3.5A), minima A4 represents a folded conformation of the Alpha1 peptide, in which the Alpha1 peptide is folded into a well-defined helix-turn-helix conformation, with the oleate molecules observed to be forming a lipid core. Minima A3 and A5 shows representative structures characterized by a single folded alpha helical segment with a disordered or partially folded region, respectively. The representative structure of minima A1 is characterized by short helical segments throughout Alpha1 peptide. Minima A1, A3 and A5 could possibly represent conformational states that precede the formation of the folded state A4. Other minima observed include Minima A2, whose representative structure is characterized by a random coil conformation. Also, in Minima A6, the representative structure is characterized by a beta structure, conformationally different and far away in principal component subspace from regions with helical conformations (Fig. 3.5A). The wide basin and multiple shallow minima of the Alpha1-oleate containing system free energy surface is indicative of conformational diversity and suggests the possibility of interconversion between unfolded (Minima A2), partial helical (Minima A1, A3 and A5), and folded helical (Minima A4) conformations (Fig. 3.5A).

On the other hand, the free energy landscape of the peptide in its apo form (Fig. 3.5B) is characterized by local minima with representative conformations of poorly defined secondary structure (Minima B3, B4, B5 and B6) and minima characterized by representative structures with various partially folded helix-turn conformations (Minima B1 and B2). Compared with the Alpha1-oleate containing system (Fig. 3.5A), the Alpha1 peptide in its apo form adopts conformations with greater conformational diversity and disorder (Fig. 3.5B, Minima B3, B4, B5, B6).

A striking observation is that representative conformation of Minima A4 in the Alpha1-oleate containing system (Fig. 3.5A), which represents the folded state, cannot be observed in the free energy surface of the Alpha1 peptide in its apo form (Fig. 3.5B). It is noted that short helical segments may form in the apo system where oleate is absent, as seen for representative structures of minima B1 and B2 (Fig. 3.5B). However, the two defined helical segments as observed in minima A4 of the Alpha1-oleate containing system (Fig. 3.5A) could not be observed in any minima of the Alpha1 peptide in its apo form (Fig.3.5B). While minima B2 in the free energy surface of the Alpha1 peptide in its Apo form lies close in Principal Component subspace to minima A4, it is observed that the two alpha helical segments in minima B2 were not as completely folded (Fig 3.5A, Fig. 3.5B). This comparison can be made because the dihedral angles of both systems were combined during the generation of the covariance matrix. The presence of oleate molecules induces a change in the free energy surface of the Alpha1 peptide. Based on these observations (Fig. 3.5A., Fig. 3.5B.), it is hypothesized that the presence of the oleate molecules serve to stabilize the helical conformations of the Alpha1-peptide. It is also observed in the representative structures of Minima, A3, A4, and A5 that in the presence of well folded helical segments, the oleate molecules are bound in close proximity (Fig. 3.5A). Conversely, it is observed that in minima A1, A2 and A6, where representative conformations lack well defined alpha helical segments, the oleate molecules are far apart and not interacting with the Alpha1 peptide (Fig. 3.5A).

3.3.4 Geometric clustering of conformations for both systems confirms findings of Dihedral Angle Principal Component Analysis

It is important to interpret the results of DPCA projection very carefully. For example, one potential effect of DPCA is the possibility of projecting multiple conformations into one minima. Thus, we sought to confirm our findings from the previous section through geometric

clustering of conformations for both the Alpha1-oleate containing system and the Alpha1 peptide in its apo form. Geometric clustering was performed using the Gromos algorithm with a cutoff of 0.55nm of the last 900ns with every 10ps of frames. Representative structures of conformations representing more than 5% of the population are shown. It is observed that conformations representing Alpha1-oleate containing system shows representative conformations containing well defined helical segments of the helix-turn-helix mortif (Fig. 3.6A). However, the representative conformations of the Alpha1 peptide in its apo form adopts a wide diversity of conformations comprising partial helix-turn and random coil conformations (Fig. 3.6B). The findings of geometric clustering (Fig. 3.6) are consistent with the DPCA free energy surface projections (Fig. 3.5), whereby similar conformations are observed for each system using both methods. The Geometric clustering results further emphasizes the conformation differences between the Alpha1 peptide in the presence and absence of oleate (Fig. 3.6).

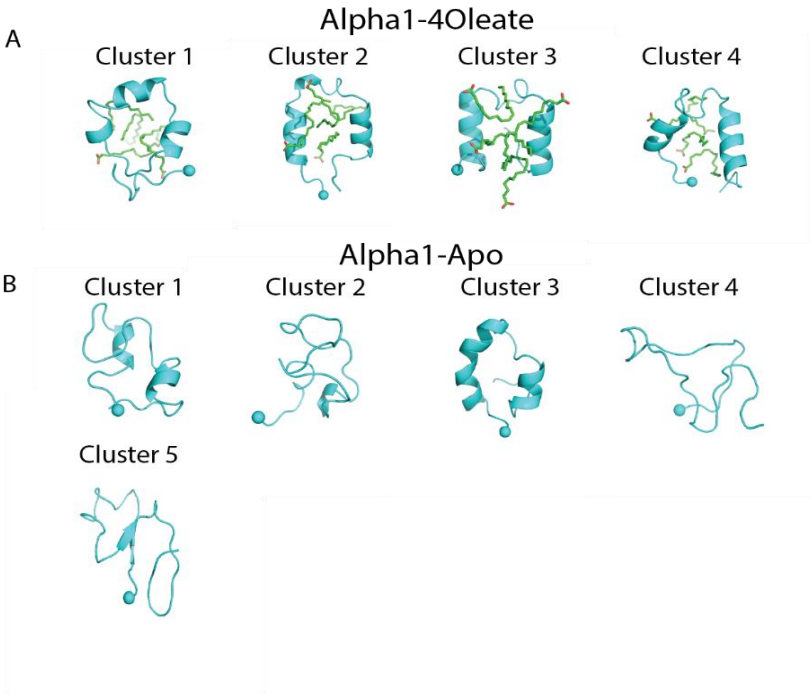

Figure 3.6. Representative conformations from geometric clustering of HAMLET Alpha1 peptide for the oleate containing system (A) and apo system (B). Alpha1 peptide shown as cyan cartoon. Oleate molecules shown as green sticks. The N-Terminus of the Alpha1 peptide is shown as a cyan sphere.

3.3.5 Contact probability calculations between Alpha1 and oleate molecules reveal high contact probabilities between hydrophobic residues and oleate molecules

We next sought to characterize the interactions between the Alpha1 peptide and oleate molecules. Our experimental collaborators used Nuclear Overhauser Effect (NOE) Spectroscopy (NOESY) and identified NOE signals between the protons in Alpha1 aliphatic side chain residues and the olefinic (Protons of the oleate molecules) protons, as well as the aromatic ring protons of the Alpha1 peptide and the olefinic protons (NMR experiments performed by experimental collaborators). The NOE is a phenomenon commonly exploited in NMR spectroscopy to provide distance information between nuclei. The NOE describes non-covalent, through-space interactions, and can be used to generate pair-wise distance restraints by a relation to a $1/r^6$ distance function[119]. However, the conformational flexibility of a system makes the resolving and interpretation of spectra signals challenging[119], and in this project specific residue-pairwise restraints information could not be obtained. Based on these findings, we calculated the contact probability between side chain protons of residues for the Alpha1 peptide and the olefinic protons. Firstly, we calculated the minimum distances between side chain protons of residues for the Alpha1 peptide and the olefinic protons. To calculate contact probability, a contact was defined if the minimum calculated distances was less than 0.45nm, and contact probability counted over the total number of frames sampled. Conventionally, contact probability calculations during simulation are performed using heavy atoms of a simulation system. However, in this case, our aim was to make comparison to NMR experiments, in which signals observed were attributed to hydrogen nuclei. Indeed, we have identified high contact probabilities between aliphatic and aromatic residues protons towards olefinic protons. High contact probabilities of more than 0.5 are identified between side chain protons of Leu8, Leu11, Leu12, Ile 15, Tyr18, Ile21, Leu23, Pro24, Leu26, Ile27, Met30, and Phe31, and that of olefinic protons (Fig. 3.7). The contact probability calculations show that our simulation conformation space is in agreement with NMR spectroscopy results. In this

study, the NMR spectroscopy measurements also confirms the close proximity of oleate molecules to the Alpha1 peptide. Such conformations in which the Alpha1 peptide is in close proximity (0.45nm cutoff) of oleate molecules may be observed in our simulation generated ensembles (Fig. 3.5A).

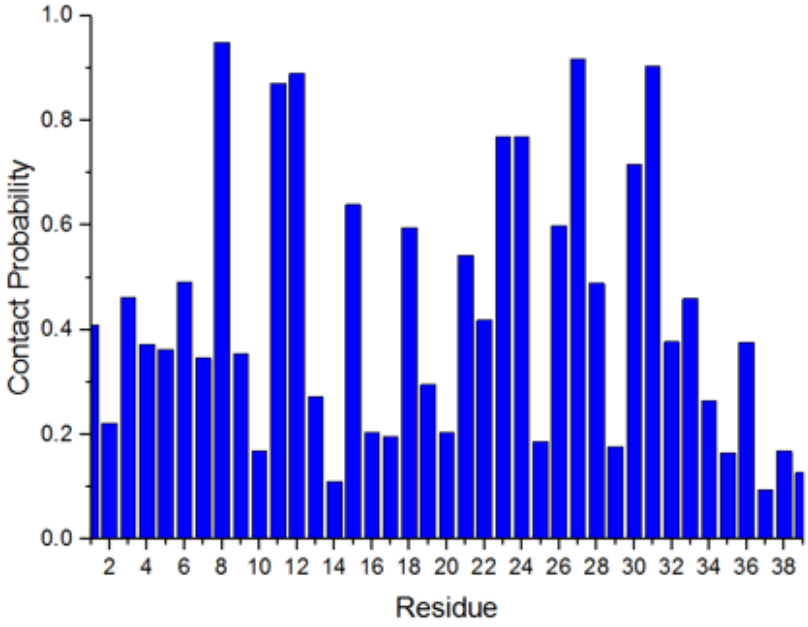

Figure 3.7. Contact probabilities between Alpha1 peptide residue sidechain protons and olefinic protons of any oleate molecule. A contact was defined if the minimum distance between side chain protons of residues from the Alpha1 peptide and olefinic protons was less than 0.45nm, and contact probability was calculated over the total number of simulation frames sampled in the last 900ns of H-REMD simulation.

3.3.6 Contact Probability analysis between Alpha1 peptide and Oleate in folded state

Our findings in the previous section showing high contact probabilities between the hydrophobic side chain residues and oleate olefinic protons motivated us to further characterize the interactions between oleate and the Alpha1 peptide in the folded conformation at atomistic detail. To do so, we extracted conformations of the folded state in the Alpha1-oleate containing system, corresponding to cluster 2 and cluster 3 from the geometric clustering (Fig. 3.6A). Following which, we performed contact probability calculations between heavy atoms of each

oleate molecule and residues of the Alpha1 peptide. Firstly, we calculated the minimum distances between the heavy atoms of oleate molecules and heavy atoms of each residue in the Alpha1 peptide. Each oleate molecule is assigned an arbitrary residue index from oleate1 to oleate4. A contact is defined if the minimum calculated distance is less than 0.45nm, and contact probability calculated over the total number of frames. To keep the dimension of the contact map low and yet provide a meaningful analysis, contact probability calculations were performed for alternate heavy atoms of each oleate molecule (ie. C1, C3, C5, C7… etc). The contact map of contact probabilities between the Alpha1 peptide and oleate molecules is shown in Figure 3.8. The contact probability map shows the extreme conformational diversity of the interactions between the Alpha1 peptide and oleate molecules in the folded state. It shows numerous contact sites without specific regions of dominantly high contact probabilities. Contact probabilities of more than 0.03 are shown, and the highest contact probability observed in the contact map is 0.11 (Fig. 3.8). This finding indicates that the binding between the Alpha1 peptide and oleate does not show any apparent specific pattern (Fig.3.8). Such non-specific binding between the Alpha1 peptide and oleate molecules could be attributed to the non-specific nature of hydrophobic interactions. The interaction between the Alpha1 peptide and oleate molecules are observed to be dominantly due to hydrophobic interactions, given the observed contacts of the Alpha1 peptide with aliphatic atoms of the oleate carbon chain (Fig. 3.8). Additionally, electrostatic interactions may be observed between the N terminal Lysine residues of the Alpha1 peptide and carboxyl oxygens of oleate molecules (Fig. 3.8). The significance of such electrostatic interactions could be further investigated experimentally.

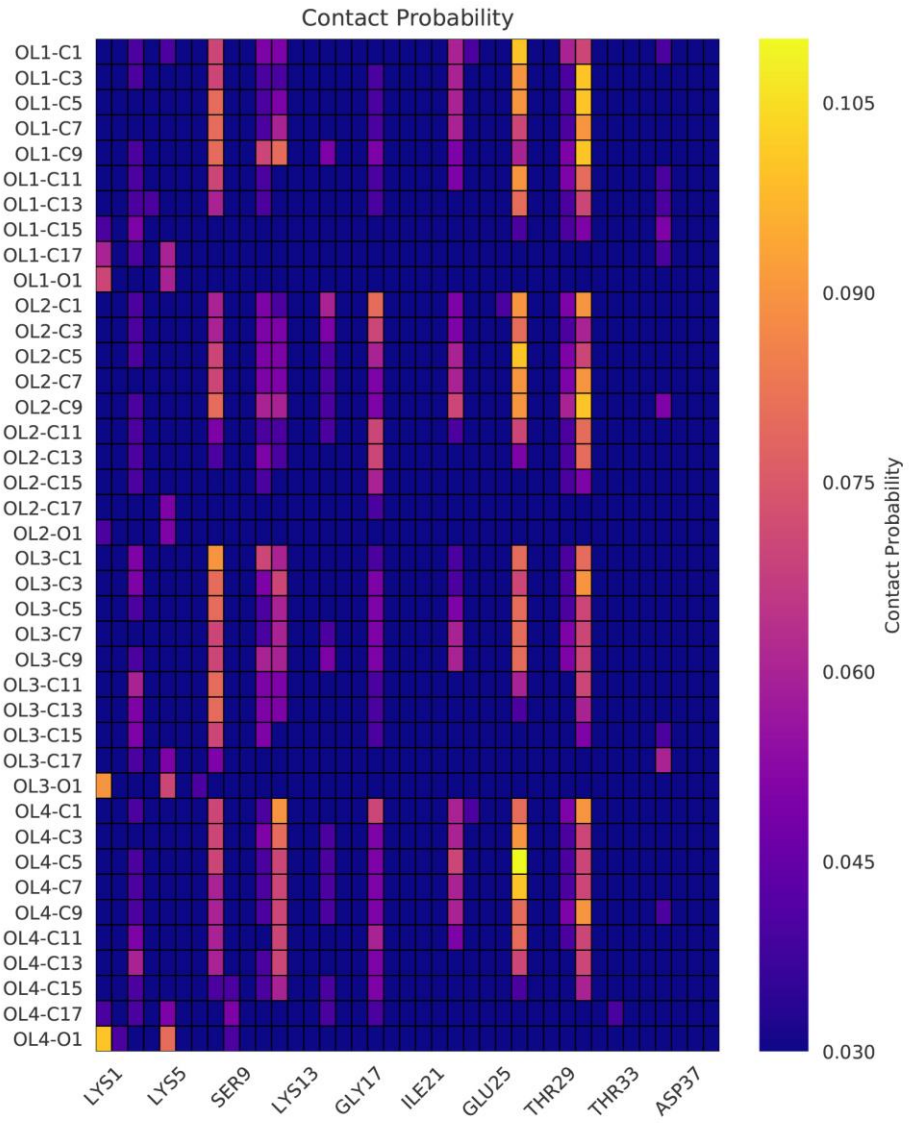

Figure 3.8 Contact probabilities between heavy atoms of Alpha1 peptide residues and heavy atoms of individual oleate molecules in the folded state. Each oleate molecule is assigned a label oleate1-4. For contact probability calculation, a contact is defined if the minimum distance between the heavy atoms of oleate molecules and heavy atoms of each residue in the Alpha1 peptide is less than 0.45nm. Contact probabilities of more than 0.03 are shown.

3.3.7 Perspective of simulation findings

Classically, it is believed that a protein chain folds into its native state, which refers to the conformation at the global free energy minimum that is associated with biological function[9]. Proteins that misfold are associated with being non-functional or deposits to form harmful aggregates, such as in the case of the amyloid-beta peptide and its correlated presence in Alzheimer's disease [13]. The discovery of HAMLET represents a novel and different category of protein function. Human Alpha-lactalbumin, in addition to its biological function in the native state of lactose synthesis, gains a novel function of tumoricidal activity upon partial unfolding and binding with oleate cofactor[21].

From a biology perspective it is hypothesized that the existence of such proteins may be present to allow alteration of their conformational folds and binding with target partners to result in different and novel biological functions. These could potentially serve as a mechanism other than alternative splicing or post translational modifications to expand the array of protein functions available[21]. The observed free energy surface of the Alpha1 peptide is likened to that of the free energy surface of an IDP. IDPs are characterized by great conformational flexibility and the ability to adopt a wide variety of functions[16]. Indeed, the conformational heterogeneity as seen from our free energy surface projections possibly explains (Ho et al, submitted) the effect of HAMLET on multiple diverse biological targets (section 3.1.2 Biological effects of HAMLET). However, distinct from IDPs, the HAMLET Alpha1 peptide require the presence of helical mortifs and binding of oleate molecules as a cofactor, resulting in a different conformation and acquisition a new, beneficial function (Ho et al, submitted).

3.4. Conclusions and future directions

In this study, we have characterized the conformation differences between the Alpha1 peptide in the presence and absence of oleate through free energy surface projections as a function of the first 2 dihedral principal components, and geometric clustering. The conformational heterogeneity of the Alpha1 peptide is demonstrated. A major characteristic of the Alpha1 peptide in the presence of oleate molecules is the presence of defined helical mortifs and the close proximity of oleate molecules to the Alpha1 peptide. Such observations could not be observed for the Alpha1 peptide in its apo form, whose conformations are characterized by partially folded short helical segments and diverse, disordered conformations. The presence of oleic acid causes a change in the folding free energy landscape of folding. It is hypothesized that the oleate serves to stabilize helical conformations of the Alpha1 peptide.

Several future directions could be considered for the HAMLET project. Firstly, it is observed that free energy surface of the Alpha1-oleate containing system can be described by the presence of a wide basin and multiple shallow minima. The nature of such a landscape suggests the ability to interconvert between conformations[17]. Understanding the kinetics of such interconversion is worthy of a future study. Kinetics of folding can be studied computationally through Markov State Models[120]. Briefly, the various states adopted by a biological system is defined, and multiple short simulation trajectories are initiated from each state. Following which the transitions between each state are counted, which eventually provides a description of the kinetics of the biological process[121, 122]. Such an effort is computationally expensive, requiring extensive conformational sampling. This is made possible through distributed computing such as the Folding@Home project[123], or exclusive use of Supercomputing clusters. Moreover, the conformational diversity of the Alpha1 peptide and the presence of 4 oleate molecules increases the complexity of sampling. Another possible direction is to

investigate the effect of the HAMLET Alpha1 peptide-oleate complex on membranes. Such a study could provide insights as to why HAMLET affects cancer cells but not healthy, differentiated cells[19]. Our experimental collaborators are in the process of performing lipidomic studies to characterize the membrane composition differences between cancer cells and healthy cells. The availability of such information could be incorporated into molecular modelling. The studying of the direct biomolecular interaction between HAMLET with its cellular targets is a challenging approach, since the response triggered by HAMLET appears to be broad spectrum and a systems level[19].

4. Atomistic study of binding between Lipocalin type Prostaglandin D synthase (L-PGDS) and Amyloid Beta peptide(1-40) Aβ(1-40)- An in-silico study

4.1 Introduction

4.1.1 Alzheimer's Disease is a "Protein Misfolding disease"

Alzheimer's Disease (AD) is a neurodegenerative disease characterized by clinical symptoms such as the progressive loss of cognitive functions and memory loss [124]. The prevalence of AD causes a significant impact on public health, being a major cause of mortality and morbidity[125]. AD is identified by the World Health Organization to be a major public health concern, with worldwide prevalence estimated to double every 20 years [126]. In addition to the presence of neurofibrillary tangles of Tau protein, a key pathological feature of AD is the presence of amyloid plaque deposits in the brain consisting predominantly of the Amyloid beta peptide (Aβ) [127]. The Aβ peptide is generated through the cleavage of the Amyloid Precursor Protein by enzymes ß and γ secretase[128] . Various forms of Aβ are produced, with Aβ(1-40) being the most predominantly populated form and the more hydrophobic and aggregation-prone Aβ(1-42) being less abundant [129]. The deposition and aggregation of soluble Aβ monomers peptides results in the formation of oligomers and insoluble Aβ fibril aggregates that could in turn disrupt tissues membranes, leading to a loss of biological function. This is known as the Amyloid Cascade hypothesis [130], and is considered a key mechanism[131] in the pathogenesis of AD.

To this end, molecular chaperones play critical roles in disease prevention of "protein misfolding diseases", serving a neuroprotection function. They binding to mis-folded proteins in order to prevent their deposition and aggregation ensuring that they are kept in a soluble state[132]. In addition, some chaperones may also assist in refolding. Chaperones form part of a "Protein Quality Control" [133] system, and neurodegenerative diseases can arise when the

function of chaperone proteins are compromised, leading to the accumulation and deposition of mis-folded proteins[134]. It is noted that intracellular chaperones and protein degradation machinery degrades the misfolded proteins would be ineffective against the large extracellular amyloid plaques[135].

4.1.2 Lipocalin type Prostaglandin D synthase (L-PGDS) is a major amyloid beta (Aβ) chaperone

The protein Lipocalin type Prostaglandin D synthase (L-PGDS) was identified to be a major amyloid beta Aβ chaperone**,** playing a role in inhibiting Aβ aggregation. This is in addition to its natural catalytic function of catalyzing the conversion of Prostaglandin H2 to Prostaglandin D2, through the catalytic residue cysteine 65[136]. It is also known to serve as a binder and transporter of lipophilic compounds [137] due to its large cavity. L-PGDS is the second most abundant human cerebro-spinal fluid (CSF) protein, second only to Albumin [138]. L-PGDS levels are found to be decreased in the human CSF of AD patients compared to healthy patients[139]. Thioflavin T assay to investigate Aβ aggregation revealed an approximately 50% reduction of Aβ fibrils in the presence of L-PGDS compared to the control in the absence of L-PGDS [136]. In addition, in L-PGDS knockout mice, there was observed accelerated Aβ fibril deposition[136]. Taken together, these findings suggest that L-PGDS can prevent the deposition of Aβ and the aggregation of Aβ into fibrils.

Structures of L-PGDS has been studied in its apo form or in complex with ligands. The structure of mouse L-PGDS was solved using X-ray crystallography, demonstrating the Lipocalin fold, composing of beta strands forming a beta barrel, providing a large cavity in between for ligand binding[140]. The structure of catalytic mutant Human L-PGDS in complex with fatty acids was also solved, showing binding of 2 lipid molecules in the cavity[141]. The

structure of wild type L-PGDS in complex with L-PGDS substrate analog U44069, and in its apo form was solved in 2013[142].

Recently, our experimental collaborators attempted to characterize the interactions between L-PGDS and Aβ(1-40) (Aβ40). Consistent with previous findings, Thioflavin T Assays measurements shows that in the presence of L-PGDS, Aβ aggregation and fibril elongation is reduced. Interestingly, it is also discovered that L-PGDS contains dis-aggregase activity, with the ability to solubilize pre-formed amyloid fibrils. L-PGDS was shown to form a complex with Aβ40 monomers in the ratio of 1:1. Using 1H-15N 2D Heteronuclear Single Quantum Coherence (HSQC) NMR spectroscopy, they attempted to identify residues with perturbations in chemical shifts upon titration of isotope labelled Aβ40 to L-PGDS, and vice versa. However, because of extensive line broadening and peak disappearance associated with conformational exchange, the sensitivity and resolution of the spectra is affected, making spectral assignment and characterization of ligand binding challenging. These findings motivated us to study the interactions between L-PGDS and Aβ40.

4.1.3 Aim of project and scope

In this project, we aim to construct a model detailing the interactions between L-PGDS and Aβ40 peptide. This model would provide insights into the mechanism of biomolecular interactions between Aβ40 and L-PGDS, thus providing an understanding of the Aβ chaperone activity of L-PGDS. This project provides the first model detailing the interactions between L-PGDS and Aβ40.

Since specific residue pair-wise restraints between L-PGDS and Aβ40 could not be obtained from NMR spectroscopy, we adopted a biophysics-based approach towards model construction. We used enhanced sampling and subsequently extracted high probability conformations of the L-PGDS- Aβ40 complex based on our sampling. We then show that our simulation generated

model agrees with experimental data. To do so, we constructed a model of L-PGDS in complex with Aβ40 by using the initial coordinates of Aβ peptide from an available crystal structure of Anticalin US7 in complex with Aβ40, of which residues 16 to 28 of Aβ peptide was resolved[143]. Anticalin US7 was generated by performing affinity maturation of Lipocalin 2 towards Aβ peptide[143]. We performed sequence alignment using Uniprot and found 37% sequence identity between Lipocalin 2 and L-PGDS (Fig. 4.1).

```
P41222 PTGDS_HUMAN          1  MATHHTLWMGLALLGVLGDL---------QAAPEAQVSVQPNFQQDKFLGRWFSAGLASN       51
P80188 NGAL_HUMAN           1  -MPLGLLWLGLALLGALHAQAQDSTSDLIPAPPLSKVPLQQNFQDNQFQGKWYVVGLAGN       59
                                    **:******.*           * * ::* :* ***:::* *:*: .***.*

P41222 PTGDS_HUMAN         52  SSWLREKKAALSMCKSVVAPATDGGLNLTSTFLRKNQCETRTMLLQPAGSLGSYSYRS--      109
P80188 NGAL_HUMAN          60  AI-LREDKDPQKMYATIYELKEDKSYNVTSVLFRKKKCDYWIRTFVPGCQPGEFTLGNIK      118
                               :  ***.*   .*  ::    * . *:**.::**::*:    : *. . *.::  .

P41222 PTGDS_HUMAN        110  -PHWGSTYSVSVVETDYDQYALLYSQGSKGPGEDFRMATLYSRTQTPRAELKEKFTAFCK      168
P80188 NGAL_HUMAN         119  SYPGLTSYLVRVVSTNYNQHAMVFFKKVSQN-REYFKITLYGRTKELTSELKENFIRFSK      177
                                   ::* * **.*:*:*:*::: :  .   .::   ***.**:   :****:*  *.*

P41222 PTGDS_HUMAN        169  AQGFTEDTIVFLPQTDKCMTEQ                                      190
P80188 NGAL_HUMAN         178  SLGLPENHIVFPVPIDQCIDG-                                      198
                               : *: *: ***    *:*:
```

Figure 4.1 Sequence alignment between human L-PGDS (PTGDS Human) (P41222) and Lipocalin 2 (NGAL human) (P80188). Sequence alignment performed on uniport. Asterisk (*) indicates identical residues, while colon (:) indicates residues with similar chemical properties.

Moreover, the structure of Anticalin US7 contains a similar structural lipocalin fold as that of L-PGDS, and it could be assumed that Aβ would bind similarly to L-PGDS as that of Anticalin US7. We have superimposed the crystal structure of Anticalin US7 in complex with Aβ(16-28) (PDB ID: 4MVI) resolved onto the crystal structure of L-PGDS in apo form (PDB ID:4IMN) for visualization in figure 4.2.

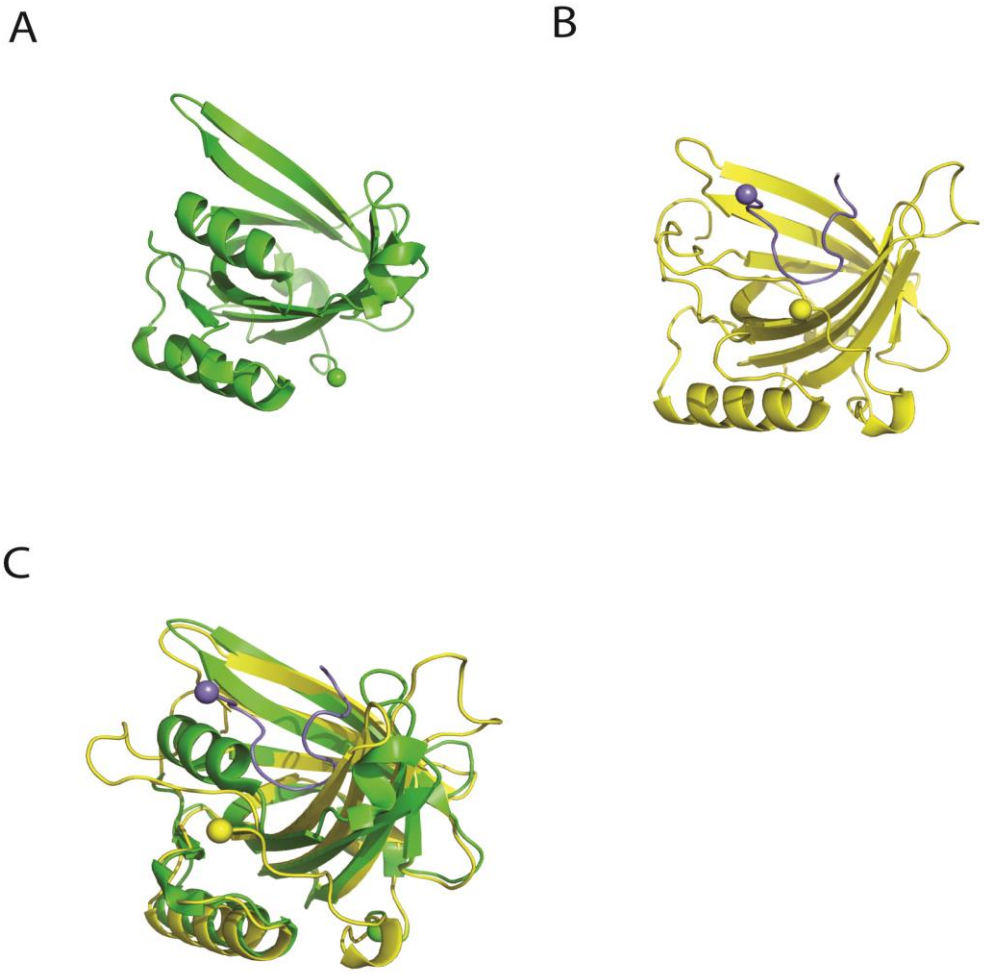

Figure 4.2. Visualization of crystal structures of L-PGDS and Anticalin US7. A. Crystal structure of L-PGDS in its apo form (PDB ID 4IMN). B. Crystal structure of Anticalin US7 with Aβ (16-28) resolved (PDB ID 4MVI). C. Superimposition of Anticalin US7 structure with Aβ (16-28) resolved onto the structure of L-PGDS. L-PGDS is shown as green cartoon. Anticalin US7 is shown as yellow cartoon. Aβ peptide shown as purple cartoon, with N-terminus shown as a purple sphere.

We performed Hamiltonian Replica Exchange Molecular Dynamics (H-REMD) using the Replica Exchange with Solute Scaling method (REST2)[64] through scaling residues of Aβ40 peptide in order to sample the conformation space of Aβ40 in complex with L-PGDS. Following which, binding modes of L-PGDS towards Aβ40 were extracted using a clustering method based on high contact probability residue pairs[144] between L-PGDS and Aβ40. The

extracted binding modes were subjected to classical molecular dynamics simulations and further analyses in order to evaluate the relative stabilities of each binding mode and to characterize biomolecular interactions between L-PGDS and Aβ40. The contact-probability based clustering method which we adopted in this study makes use of sampling statistics and selecting conformations with high contact probabilities. In contrast, a molecular docking approach involves predicting the binding pose of a ligand using a search algorithm and evaluating the pose using a scoring function. The molecular docking approach while requiring relatively lower computational resources to perform would result in a compromise in accuracy[145].

4.2 Method

4.2.1 Initial construction of L-PGDS- Aβ40 model

The initial structure of L-PGDS in its apo form was taken from Lim et al, 2013 (PDB ID: 4IMN). The structure of Anticalin US7 in complex with Amyloid Peptide Aβ40 (PDB ID: 4MVI)[143] contained residues 16-28 of Aβ40 resolved, and was used to construct the initial model of L-PGDS in complex with Aβ40. The crystal structure of Anticalin US7 in complex with Aβ (16-28) (PDB ID 4MVI) was superimposed onto to the structure of L-PGDS (PDB ID: 4IMN). Following which, coordinates of L-PGDS (from PDB ID: 4IMN) and Aβ(16-28) (from PDB ID: 4MVI) were extracted to generate the initial model of L-PGDS with Aβ(16-28). The unresolved residues of Aβ40, ie. residue D1-Q15 and G27 to V40, were modelled in an arbitrary extended manner using Discovery Studio 4.1[107]. The L-PGDS- Aβ40 complex was subjected to a short steepest descent energy minimization in-vacuo using the Gromacs[58] 5.1.2 package and the charmm36M[56] force field, until a force convergence of 1000.0 kJ/mol/nm is achieved.

4.2.2 Hamiltonian Replica Exchange Molecular Dynamics

In order to study the Binding Modes of Aβ40 in the complex L-PGDS, we performed H-REMD simulation of the L-PGDS-Aβ40 complex. The Replica Exchange with Solute Scaling (REST2)[64] method was used, and simulations were performed using the Gromacs[58] 5.1.2 package with the Plumed[114] 2.3 plugin. The CHARMM36m force field was used, whereby the parameter sets have been optimized for simulations of folded and Intrinsically Disordered Proteins [56]. The residues of Aβ40 peptide were selected for Hamiltonian Scaling. Sixteen replicas and an effective temperature range of 300K-600K following a geometric progression were used to generate the Hamiltonian scaling factors. The temperatures are 300, 314.188, 329.047, 344.61, 360.908, 377.976, 395.852, 414.574, 434.181, 454.715, 476.22, 498.743, 522.33, 547.033, 572.905, and 600K. This resulted in an average exchange probability of approximately 30%. The system was solvated with CHARMM-modified TIP3P[105] water in a cubic box with a distance of 1.2nm from the protein to the box edge. Na+ and Cl- counterions were added to neutralize the system to a concentration of 0.15M. Bonds containing hydrogen atoms were constrained using the LINCS[111] algorithm, to enable a time step of 2fs. The Particle Mesh Ewald[112]method was implemented for electrostatics. A cutoff of 1.2nm was used for both electrostatics for van der Waal's interactions. A short steepest Descent energy minimization was performed to remove any initial bad contacts. An equilibration of 1ns in the NVT ensemble was performed prior to the start of production simulation. Temperature of the system was maintained using the V-rescale thermostat[113]. Exchanges were attempted every 2ps and the simulation was performed for 600ns.

4.2.3 Contact Probability based clustering method

Analysis of the simulation was performed on the ensemble of the unscaled potential energy (with effective temperature T=300K) from 100-600ns. In a system which is highly dynamic

and flexible, such as one containing an intrinsically disordered peptide, using conventional structure-based clustering methods would hinder the search for specific binding modes. The binding modes of Aβ40 in complex with L-PGDS were identified using a contact-probability based clustering method similar to Liu et al [144]. Briefly, residue-pairs between L-PGDS and Aβ40 with high contact probabilities are identified, and each simulation frame is searched for the presence of such contact pairs and re-labelled accordingly. To do so, we calculated the minimum distances between L-PGDS and Aβ40 residues from 100-600ns of the H-REMD simulation using the gmx mindist tool of the Gromacs[58] package. Only heavy atoms (non-hydrogen) atoms were considered in the calculation. Following which, a matrix of contact probabilities between L-PGDS and Aβ40 residues were generated. For contact probability calculation, a contact was counted if the minimum distance between residue pairs was less than 0.45nm. The threshold value of 0.45nm for the minimum distance to calculate contact probability was chosen after trying several threshold values. On one hand, if the threshold for minimum distance is too low, low contact probability values will result between residue pairs. This leads to a situation where it is difficult to determine which residue pairs are considered significant since many residue pairs low contact probabilities. On the other hand, if the minimum distance threshold is too high, high contact probability values between residue pairs will result. This leads to a situation where too many residue pairs have high contact probability values. Each simulation frame is re-labelled based on the presence of these contact pairs, since not all contacts may be involved for a specific binding mode. For example, for a particular frame labelled "ABCDE", it implies that the residue pairs A, B, C, D and E are present in close contact. A contact probability of more than 0.49 is considered to select contacts for frame relabelling. The threshold value for the contact probability of more than 0.49 to be considered as significant for frame re-labelling was determined by plotting a histogram of contact probability values for all residue pairs and observing the skew of the histogram. This value was

found to be appropriate because it resulted in 22 significant residue pairs giving a labelling coverage of 99.4%, with only 1471 out of 250001 (0.58%) frames remaining un-labelled. All frames can be classified by a specific label, and the label comprising various combinations of letters, serves as a classifier for binding modes. Subsequently, to identify the representative structure of each binding mode, we performed geometric clustering of the frames corresponding to a particular binding mode using the Gromos algorithm in the in-built Gromacs[58] gmx cluster tool. The group for the clustering was based on the close contact residue pairs with high contact probability belonging to that particular binding mode. A clustering cutoff of 0.25nm was used.

4.2.4 Classical Molecular Dynamics simulations and Analysis

To perform further analysis of the shortlisted binding modes, classical molecular dynamics simulations were initiated from the representative structures of each identified binding mode. For each selected binding mode, three repeats of 200ns classical molecular dynamics simulations were performed with differing random initial velocities. For comparison, we also performed classical Molecular Dynamics simulations of L-PGDS in its apo form using crystal structure (PDB ID: 4IMN) as the initial structure.

The biomolecular interactions for each binding mode were identified based on definitions of biomolecular interactions, as discussed in (Section 2.6 Identification of Biomolecular Interactions). Contact occupancies of each identified biomolecular interaction in each binding mode were calculated throughout three simulation repeats over the last 150ns of each simulation repeat.

Subsequent analyses were performed to evaluate the stability of each binding mode. RMSD calculations of high contact probability residue pairs were performed for each binding mode

using in-built Gromacs gmx rms tool. Also, MM-PBSA calculations were performed for the last 20ns for each binding mode, using the g_mmpbsa[146] tool. Additionally, properties such as Solvent Accessible Surface Area (SASA) and Radius of Gyration (Rg) calculations were performed with respective gmx tools.

4.3 Results and discussion

In this study, we aim to understand the atomistic interactions between L-PGDS and Aβ40. To do so, we used H-REMD to sample the conformation space of Aβ40 in complex with L-PGDS. Following which, we employed a clustering method based on high contact probability residue pairs between L-PGDS and Aβ40 to identify binding modes.

4.3.1 Convergence of H-REMD simulations

To evaluate the convergence of the H-REMD simulation, we evaluated several properties of the simulation. We plotted the diffusiveness of H-REMD replicas in effective temperature space for a representative low, medium and high effective temperature replica (Fig. 4.3A). The diffusive behaviour of H-REMD replicas was observed, suggesting replicas from higher effective temperatures were able to exchange conformations and contribute to enhance the sampling of the lower effective temperature replica with effective temperature 300K. We also plotted the minimum distances between residue heavy atoms of L-PGDS and Aβ40, which is a reaction coordinate describing the binding of L-PGDS to the Aβ peptide (Fig. 4.3B). The minimum distance can serve as a descriptor of the binding between L-PGDS and the Aβ peptide. When the minimum distance between L-PGDS and the Aβ peptide is low, it indicates that L-PGDS and the Aβ peptide are close in space. They are hence likely to form biomolecular interactions between each other, since the strength of a biomolecular interaction decreases with distance. On the other hand, when the minimum distance between L-PGDS and the Aβ peptide is high, it indicates that L-PGDS and the Aβ peptide are far apart in space. In addition, we

monitored properties of the system, such as the secondary structure changes as a function of time for coil, bend and turn secondary structures of the Aβ40 peptide (Fig. 4.3C.,D.,E.). These plotted properties do not change significantly as a function of simulation time. Taken together, these suggest that the simulation sampling achieved reasonable convergence for the conformation space we sampled.

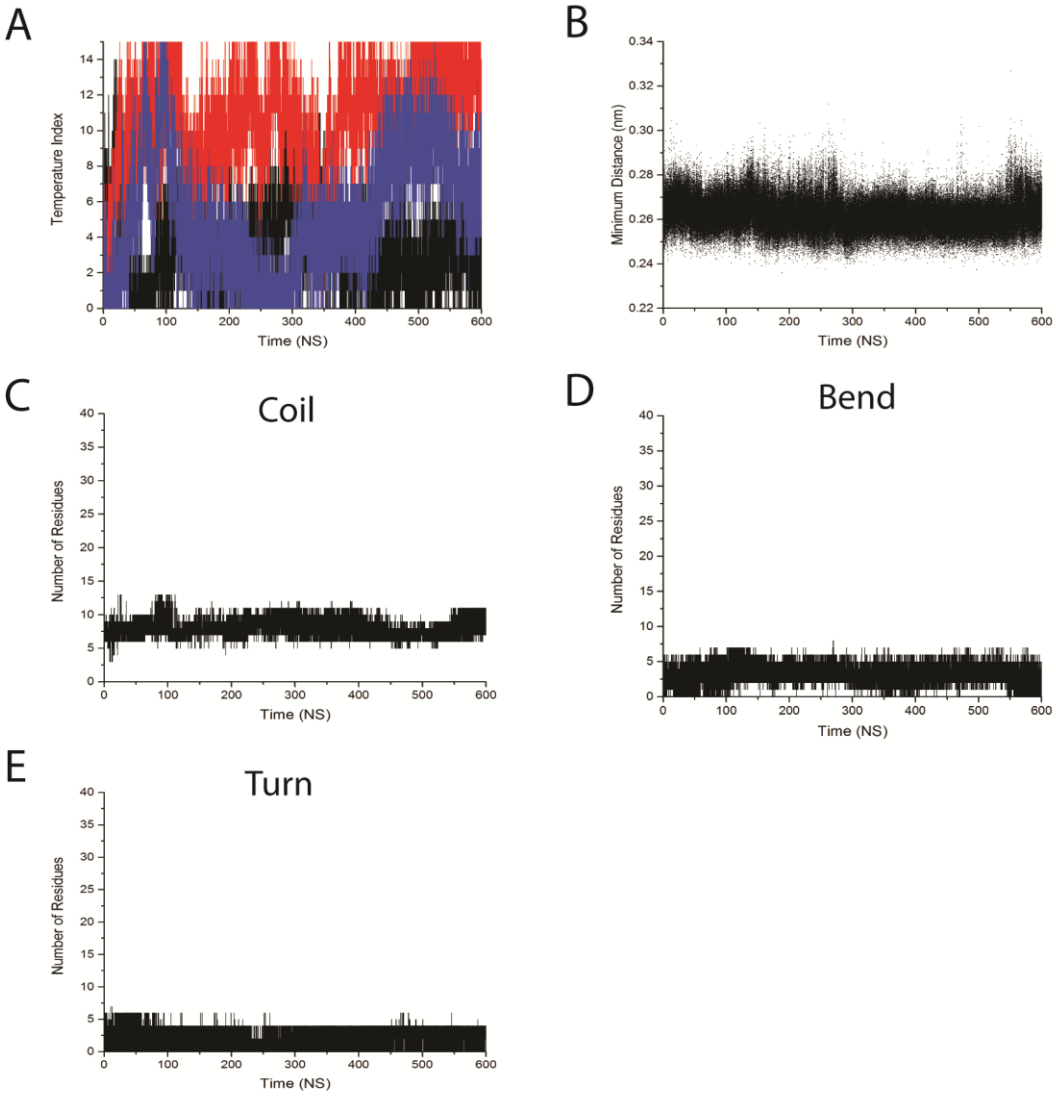

Figure 4.3 Convergence of H-REMD simulation for L-PGDS-Aβ40 complex. A. Diffusiveness of H-REMD replicas were plotted for a representative low (effective temperature= 300K, black), medium (effective temperature= 414.57K, red) and high (effective temperature= 600K, blue) effective temperature replica. B. Minimum distance between heavy atoms of L-PGDS and Aβ40. C.-E. Secondary structure propensities for Coil, Bend and Turn secondary structure, respectively as a function of simulation time.

4.3.2 Contact probability calculations reveal high contact probability residue pairs between L-PGDS and Aβ40

To understand the interactions between L-PGDS and Aβ40, we calculated the pairwise residue contact probabilities between L-PGDS and Aβ40. A contact is considered if the residue heavy atom distances between L-PGDS and Aβ40 is less than 0.45nm. Contact probability is calculated over the total number of simulation frames from 100ns-600ns. A heatmap in figure 4.4 summarizes the residue pairwise contact probabilities between L-PGDS and Aβ40, from a contact probability range of 0 (white) to 1 (dark blue). It can be observed from the heat map that there are several regions of high contact probability between L-PGDS and Aβ40 residues (Fig. 4.4). We considered a contact probability of above 0.49 to be significant, and as a result identified 22 high contact probability residue pairs between L-PGDS and Aβ40 labelled A to V, which we used to relabel all simulation frames from 100-600ns. The initial 100ns of simulations were discarded to remove the effects of initial equilibration. For clarity, the residue pair label and corresponding contact probability of each identified residue pair has been summarized in Table 4.1.

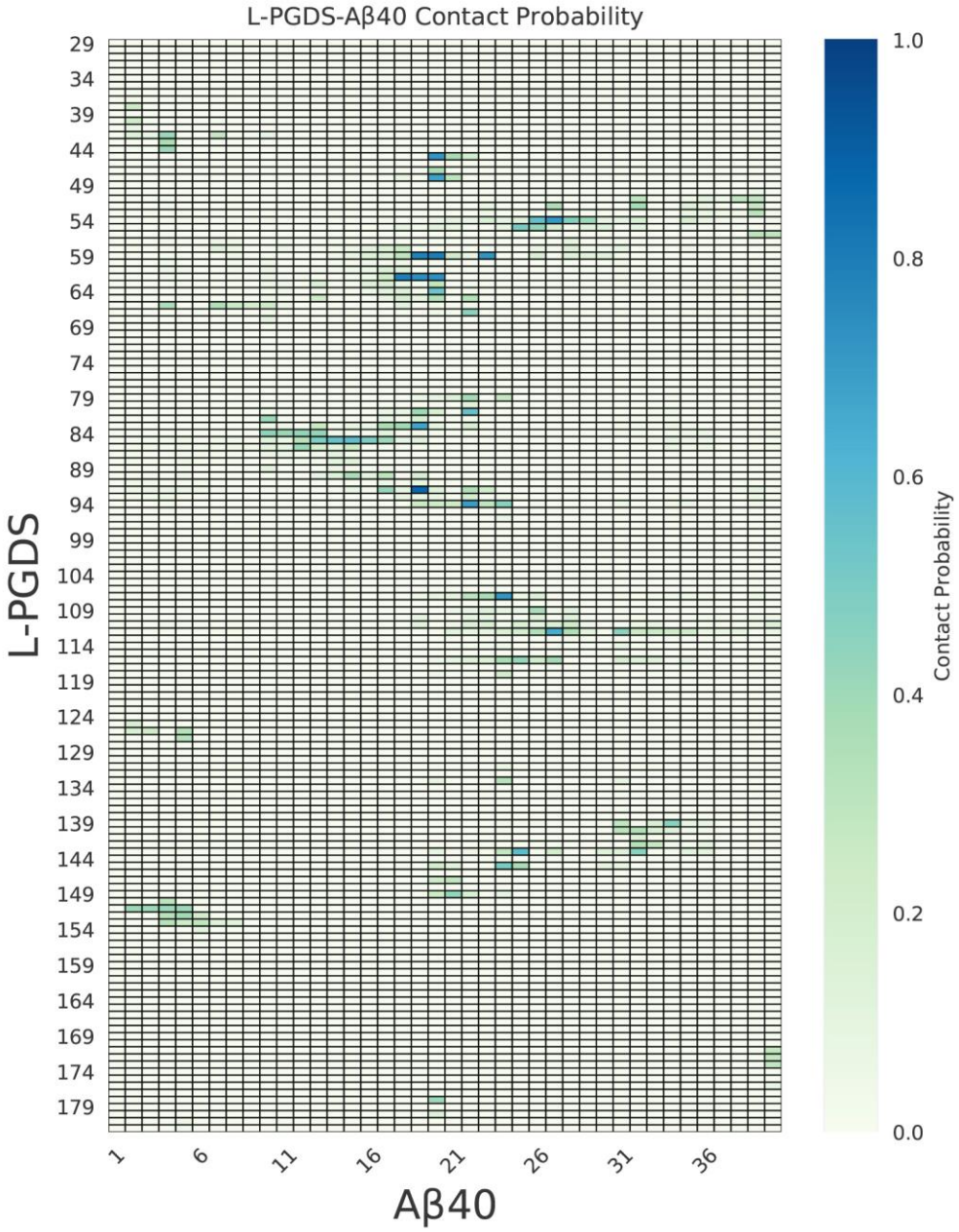

Figure 4.4 Heat map of contact probabilities between residues of L-PGDS and Aβ40. Contact probabilities were calculated based on a minimum distance cutoff of 0.45nm between heavy atoms of L-PGDS and Aβ40 residues. Contact probabilities were calculated over the total number of simulation frames from 100ns-600ns of H-REMD simulation. The contact probability ranges from 0 (white) to 1 (dark blue) is shown.

Table 4.1 High contact probability residue pairs between L-PGDS and Aβ40 with their respective contact probabilities

Label	LPGDS	Aβ40	Probability
A	92	19	0.87
B	59	20	0.79
C	62	18	0.79
D	59	19	0.78
E	62	19	0.74
F	62	20	0.73
G	59	23	0.72
H	45	20	0.71
I	107	24	0.71
J	54	27	0.70
K	94	22	0.69
L	48	20	0.67
M	83	19	0.67
N	112	27	0.64
O	143	25	0.57
P	85	15	0.57
Q	54	26	0.56
R	85	14	0.55
S	81	22	0.55
T	64	20	0.54
U	145	24	0.50
V	55	26	0.50

4.3.3 Re-labelling of simulation frames with high contact probability residue pairs identifies 3 binding modes of L-PGDS towards Aβ40

We next used the 22 high contact probability residue pairs to re-label simulation frames. Using these 22 residue pairs resulted in a labelling coverage of 99.4%, with only 1471 out of 250001 (0.58%) frames remaining unlabelled. This shows that the use of 22 high contact probability residue pairs achieves robust coverage of the simulation sampled conformation space. Binding modes covering more than 2% of the total simulation population are identified, which corresponds to the top 6 most populated binding modes (Fig. 4.5). The 6 binding modes may be further reduced to 3 binding modes, given that other binding modes differ from these 3 binding modes by only 1 or 2 high contact probability residue pairs (Fig. 4.5). It is also observed from the binding mode (BM) analysis that there is a single binding mode that dominates the population. BM1, containing the residue pairs A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V, and it's variants represents about 17.9% of the population (Fig. 4.5). On the other hand, the other binding modes, BM2, containing the residue pairs A,C,N and P and BM3, containing residue pairs A and M, both represents less than 2.5% of the population. Table 4.2 shows a summary of each of the 3 binding modes and their respective residue pairs.

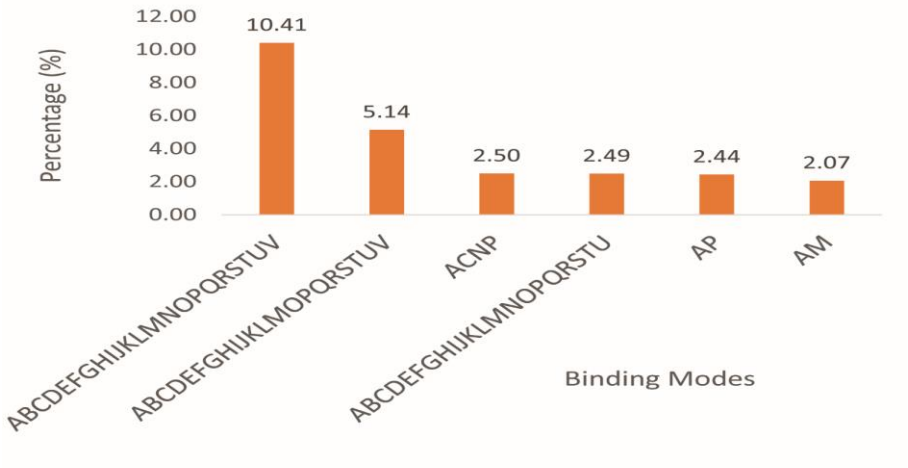

Figure 4.5. Top 6 binding modes of L-PGDS towards Aβ40 and their percentages with respect to the total simulation frames population

Table 4.2 Three identified binding modes (BM) of L-PGDS- Aβ40 interaction

Binding Mode (BM)	Label	LPGDS	Aβ40	Probability
BM1 (ABCDEFGHIJKLMNOPQRSTUV)	A	92	19	0.87
	B	59	20	0.79
	C	62	18	0.79
	D	59	19	0.78
	E	62	19	0.74
	F	62	20	0.73
	G	59	23	0.72
	H	45	20	0.71
	I	107	24	0.71
	J	54	27	0.70
	K	94	22	0.69
	L	48	20	0.67
	M	83	19	0.67
	N	112	27	0.64
	O	143	25	0.57
	P	85	15	0.57
	Q	54	26	0.56
	R	85	14	0.55
	S	81	22	0.55
	T	64	20	0.54
	U	145	24	0.50
	V	55	26	0.50
BM2 (ACNP)	A	92	19	0.87
	C	62	18	0.79
	N	112	27	0.64
	P	85	15	0.57
BM3 (AM)	A	92	19	0.87
	M	83	19	0.67

4.3.4 Biomolecular interactions between L-PGDS and Aβ40 identified through classical molecular dynamics simulations

Having identified three binding modes between L-PGDS and Aβ40, we sought next to further understand these binding modes using classical molecular dynamics simulations. To obtain the representative conformation of each binding mode, we performed geometric clustering of frames belonging to the specific binding mode. The clustering group selected was high contact probability residue pairs specific to that particular binding mode. A clustering cutoff of 0.25nm was used. Three repeats of 200ns classical molecular dynamics simulations were performed for each binding mode.

The interaction of L-PGDS towards Aβ40 constitutes a form of biomolecular recognition. This is because the binding of L-PGDS towards Aβ40 serves a biological function, which is to prevent the deposition and aggregation of Aβ40 to insoluble fibrils. We sought to identify various types of biomolecular interactions between L-PGDS and Aβ40, based on the high contact probability residue pairs between L-PGDS and Aβ40. First, we visualized the representative structures of the 3 Binding Modes. Following which, biomolecular interactions are identified based on definitions defined in Section 2.6. Identification of biomolecular interactions. For statistical significance and meaning, we also calculated the occupancies of identified biomolecular interactions over the last 150ns of three classical molecular dynamics simulations repeats for each binding mode. The representative binding poses of the three binding modes and the identified biomolecular interactions are shown in Figure 4.6. The calculated contact occupancies of each identified biomolecular interactions are shown in Table 4.3. All biomolecular interactions with more than 25% occupancy are listed.

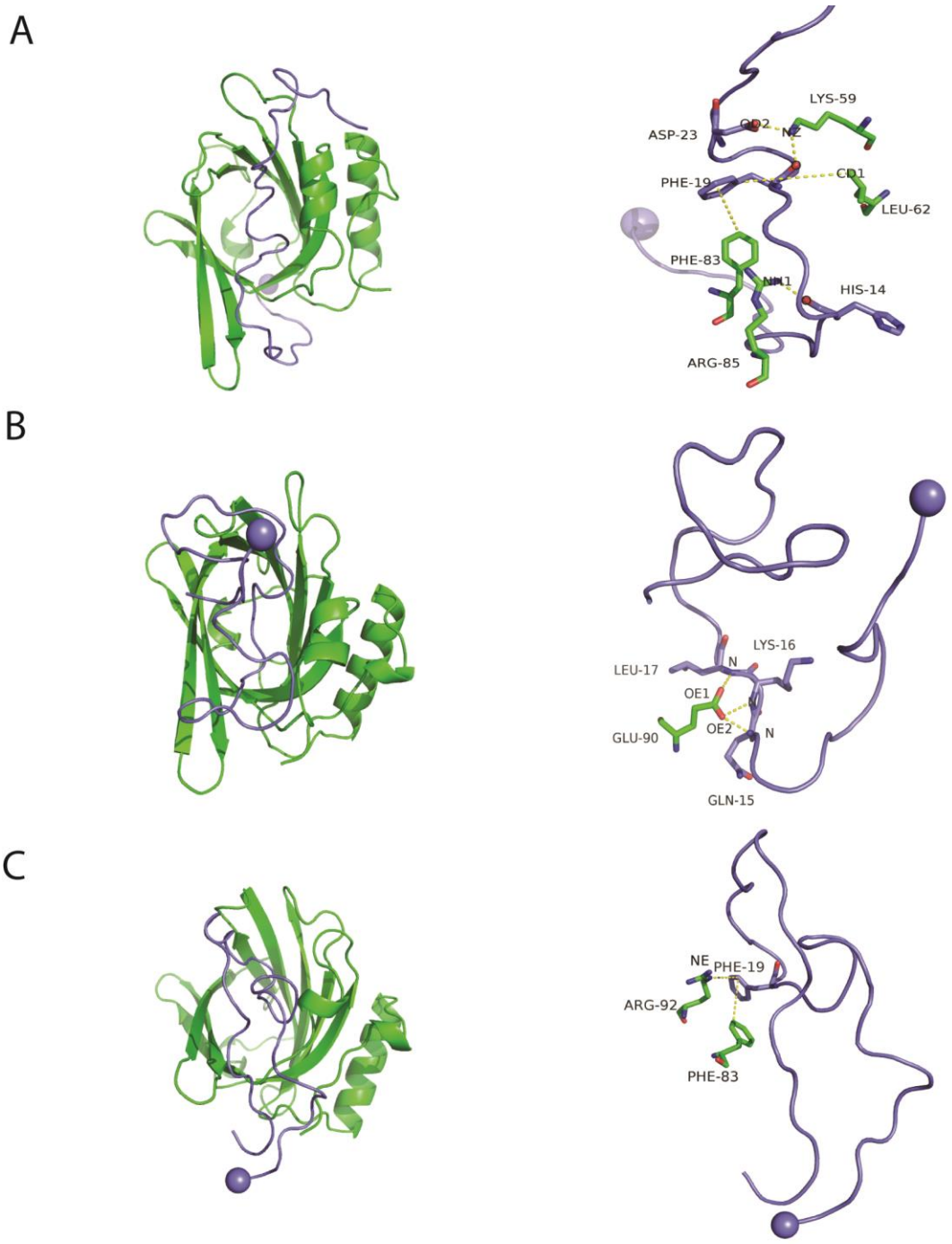

Figure 4.6 Representative binding pose of L-PGDS (Green) towards Aβ40 (Purple). A. Binding Mode 1, B. Binding Mode 2. C. Binding Mode 3. The N-Terminus of Aβ40 is shown as purple spheres. Residues identified to be involved in biomolecular recognition are shown as sticks.

Table 4.3 Table of Contact Occupancies for identified biomolecular interactions in each binding mode of L-PGDS towards Aβ40

Binding Mode	L-PGDS	Aβ40	Type	Occupancy (%)
Binding Mode 1	K59-NZ	F19-O	Hydrogen Bond	98.7
	K59-NZ	D23-OD2	Electrostatic	97.9
	L62-CD1	F19-ring	CH-π	30.4
	R85-NH1	H14-O	Hydrogen Bond	64.7
	F83-ring	F19-ring	Aromatic stacking	33.3
Binding Mode 2	E90-OE2	K16-N	Hydrogen bond	53.4
	E90-OE2	Q15-N	Hydrogen bond	47.9
	E90-OE1	L17-N	Hydrogen bond	48.2
Binding Mode 3	F83-ring	F19-ring	Aromatic Stacking	42.7
	R92-NE	F19-ring	Cation-π	27.3

Various biomolecular interactions, such as hydrogen bonds, electrostatic interactions, aromatic ring stacking, and CH-π interaction are responsible for the interactions between L-PGDS and Aβ40 (Fig. 4.6, Table. 4.3). Through visualization of biomolecular interactions and contact occupancy calculations, BM1 is also shown to contain the highest number of biomolecular interactions between L-PGDS and Aβ40 (Fig. 4.6, Table 4.3). Additionally, BM1 is observed to be the most populated binding mode among the 3 binding modes (Fig. 4.5) and contains the highest number of high contact probability residue pairs (Table 4.2). Taken together, these findings suggest that BM1 the most relatively stable binding mode of L-PGDS towards Aβ40.

4.3.5 Further Root Mean Square Deviation calculations and Binding Free energy calculations confirm the stability of Binding Mode 1

We further evaluated the stability of each binding mode by calculating the Root Mean Square Deviation (RMSD) of the high contact probability residue pairs belonging to each binding mode as a function of time. It is observed that BM1 showed the least fluctuation in RMSD as a function of simulation time (Fig. 4.7), while the RMSD values for BMs 2 and 3 fluctuated with great variability (Fig. 4.7).

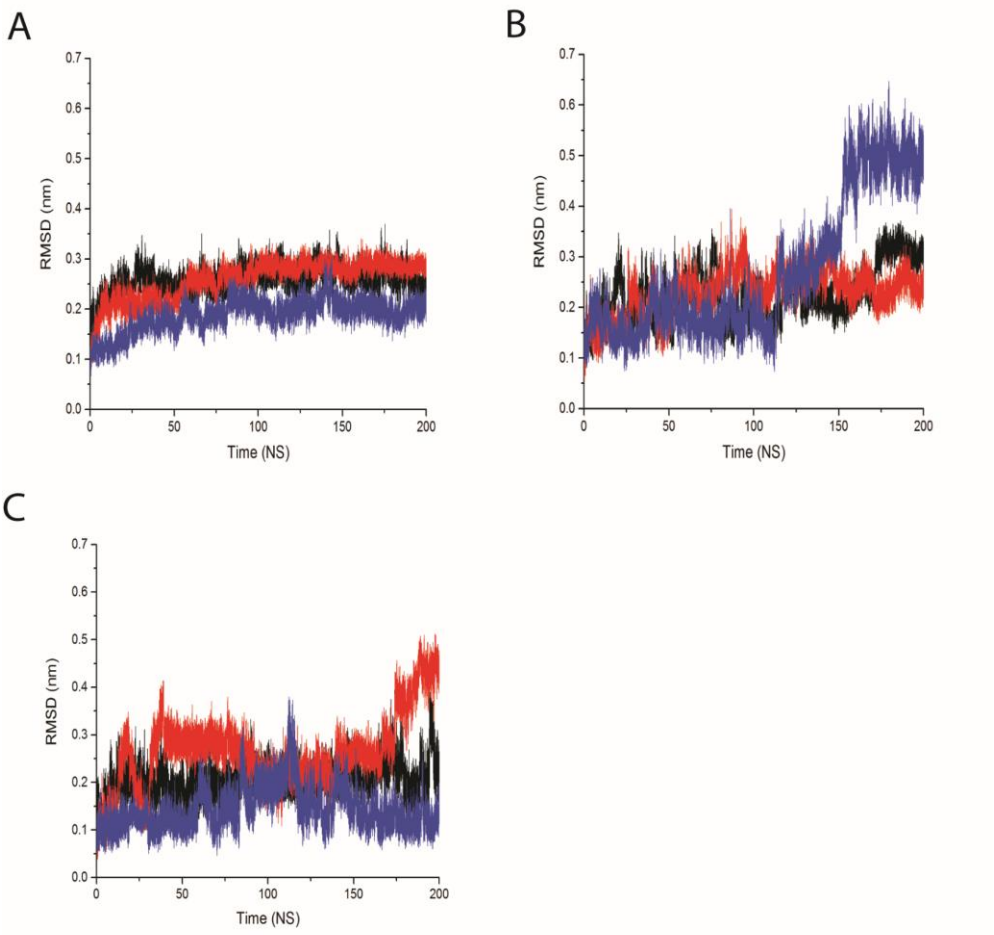

Figure 4.7 RMSD of high contact probability residue pairs as a function of time for A. Binding Mode1, B. Binding Mode2, and C. Binding Mode 3. Each of the three simulation repeats are coloured black, red and blue, respectively.

For further evaluation of the stability for each binding mode, we calculated the binding free energies of each binding mode using Molecular Mechanics Poisson Boltzmann Surface Area

(MM-PBSA) method. The last 20ns of each classical molecular dynamics simulation repeat for each binding mode was used for binding free energy calculations. The binding free energy calculation results are summarized in table 4.4.

Table 4.4 Binding Free Energy calculations for each Binding Mode (BM) of L-PGDS towards Aβ40. Energies in units of KJ/Mol.

Binding Mode	Evdw	Eelec	Epolar	Eapolar	Ebinding
BM1- ABCDEFGHIJKLMNOPQRSTUV	-760.961 ± 1.216	-1630.210 ± 6.779	1850.376 ± 5.628	-99.071 ± 0.135	-639.960 ± 2.398
BM2- ACNP	-561.760 ± 1.722	-1138.183 ± 2.415	1203.068 ± 2.564	-71.287 ± 0.203	-568.044 ± 2.266
BM3- AM	-396.525 ± 1.227	-1066.388 ± 3.717	922.888 ± 3.941	-53.425 ± 0.139	-593.700 ± 2.259

The calculated binding free energy for BM1 is the most favourable compared to BMs 2 and 3, with the most favourable (more negative) electrostatic and van der Waal's interaction energies that overcomes the free energy of solvation to allow favourable binding (Table 4.4). This is consistent with the fact that BM1 is also the binding mode with the most number of high contact probability residue pairs (Table 4.2) and biomolecular interactions (Fig. 4.6) between L-PGDS and Aβ40 compared to the other two binding modes, explaining its high electrostatic and van der Waal's interactions energies between L-PGDS and Aβ40 (Table.4.4). Note that entropy calculations are not performed for these calculations. This is because calculations of entropy is filled with high uncertainty of errors[71]. The RMSD calculations of the high contact probability residue pairs (Fig. 4.7) and the binding free energy calculations (Table. 4.4) confirm that BM1 is the most relatively stable binding modes among the 3 identified binding modes.

4.3.6 Agreement of L-PGDS and Aβ40 binding model with experimental data

For a simulation models to be meaningful, it is important that simulation generated models are in agreement with experimental data. A simulation generated model in agreement with experimental data would also be able to explain biological phenomenon and make predictions[35].

Our experimental collaborators performed NMR chemical shift perturbation titrations to understand the interactions between L-PGDS and Aβ40. Upon addition of Aβ40 peptide to 15N-labeled L-PGDS, cross-peaks stemming from residues D37, L40, S63, L77, L84, T91, A129, L130, K137, R144, M145 and F179 were shifted, while cross-peaks from residues L55, E57, T73, G76, A99, G100, S104, S114, T115, D126, and F163, showed significant attenuation in signal intensity. In addition, cross peaks from residues A49, N51, S67, L79, T82, T91, S114, V121, T123, Y128, L130, K137, G140, M145, A146, T147, K160, and I177 completely disappeared upon binding. reverse NMR titration was also performed, in which 1H-15N HSQC of 15N labelled monomeric Aβ40 peptide was recorded with and without the addition of unlabelled L-PGDS. Cross peaks from residues R5, D23, V24, L34 and M35 showed chemical shift perturbation, while residues V18, F19, F20, A21, E22, D23, V24, G25, S26, K28, G29, A30, I31, I32, G33, L34, M35, V36, G37, G38, V39 and V40 showed significant reduction in signal intensity[147], experiments performed by Ms. K. Bhuvaneswari., A/Prof Konstantin Pervushin's lab).

Chemical Shift Perturbation (CSP) arises because of changes in the nuclei environment when a ligand is titrated to a protein. These changes in chemical shifts can arise as a result of the ligand directly interacting with the protein at the interaction site[148]. To further understand the interaction between L-PGDS and Aβ40, we generated a contact map based on the average

heavy atom distances between residues of L-PGDS and Aβ40 from the last 150ns of three simulation repeats for BM1, with an average distance range of 0 (white) to beyond 1.0nm (dark blue) (Fig. 4.8). On the contact map, we were able to identify several light-coloured regions, which is indicative of L-PGDS and Aβ40 residues being in proximity of each other in our simulation (Fig. 4.8). Next, we highlighted the residues of L-PGDS which were identified by NMR to have undergone CSPs in red, and highlighted Aβ40 residues which were identified by NMR to have undergone CSPs in yellow. For clarity, CSP identified residues of L-PGDS with average simulation calculated distances of more than 1.0 nm towards all Aβ40 residues are not highlighted, and vice versa. We are able to observe intersections between residues identified by NMR to have undergone CSPs and with regions of the contact map where simulation calculated average distances between L-PGDS and Aβ40 are low (light colored regions on the contact map), thus showing that our model agrees with experimental NMR data (Fig. 4.8). Moreover, our simulation model is able to propose additional contact sites between L-PGDS and Aβ40, which are the light-coloured regions of the contact map that are in close proximity of highlighted residues (Fig. 4.8). For visualization, we have highlighted the residues identified to have undergone CSPs by NMR onto the representative structure of BM1 (Fig. 4.9). Our model shows a direct binding between L-PGDS and Aβ40 (Fig. 4.9), supported by experimental NMR measurements (Fig. 4.8).

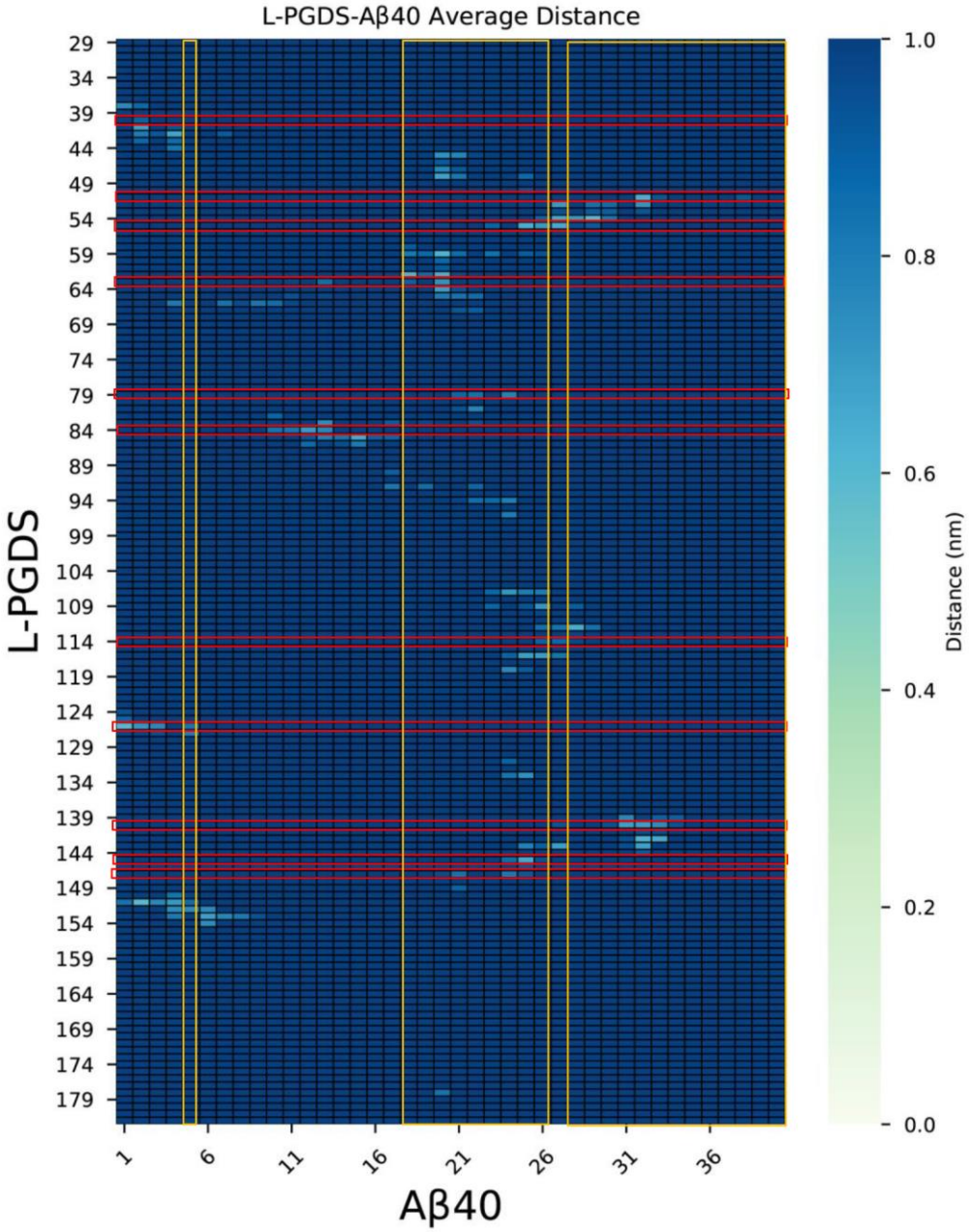

Figure 4.8. Contact map of average heavy atom distances between residues of L-PGDS and Aβ40. The average distances were calculated from the last 150ns of three simulation repeats for Binding Mode 1. The average distance was shown with a range of 0 (white) to beyond 1 nm (dark blue). Residues of L-PGDS which were identified by NMR to have undergone CSPs were highlighted in red, and Aβ40 residues which were identified by NMR to have undergone CSPs were highlighted in yellow.

It is important to note that in addition to direct binding, CSPs can also arise as result of conformational changes in the protein that were induced during ligand binding[149]. Several residues of L-PGDS in which CSPs were detected were found to be buried deep within the L-PGDS cavity of our BM1 representative model (Fig. 4.9). In an attempt to explain the presence of CSPs in these residues, we calculated the Solvent Accessible Surface Area (SASA) values per residue for each residue of the L-PGDS in apo form (Fig. 4.10). SASA values were calculated over the average of the last 150ns of three 200ns simulation repeats, and only of L-PGDS residues identified by NMR to have undergone CSPs were shown for clarity (Fig. 4.10). It is observed that several residues of L-PGDS identified by NMR to have undergone CSPs have low SASA values, thus being solvent inaccessible for direct binding (Fig. 4.10). We attribute the presence of such chemical shifts and the observed broad loss of NMR resonances to potential allosteric changes when Aβ40 binds to L-PGDS. This interesting phenomenon is worthy of further investigation in a future work. It is important to note that CSP by 2D HSQC alone cannot distinguish direct binding and allosteric effects[149]. Further NMR experiments such as NOESY, which can provide pairwise distance restraints information[119], have to be performed.

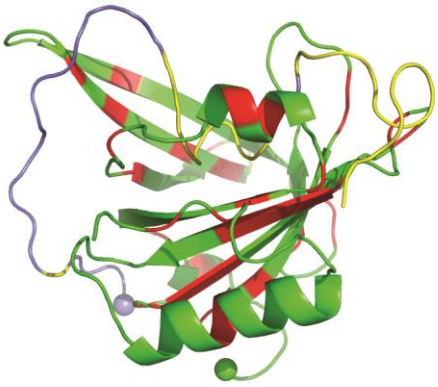

Figure 4.9. Residues experimentally identified to have undergone chemical shift perturbation through NMR were highlighted red (L-PGDS) and yellow (Aβ40). L-PGDS is shown as green cartoon. The N-Terminal of L-PGDS is shown as green spheres. Aβ40 peptide shown as purple cartoon. The N-Terminal of Aβ40 is shown as purple sphere.

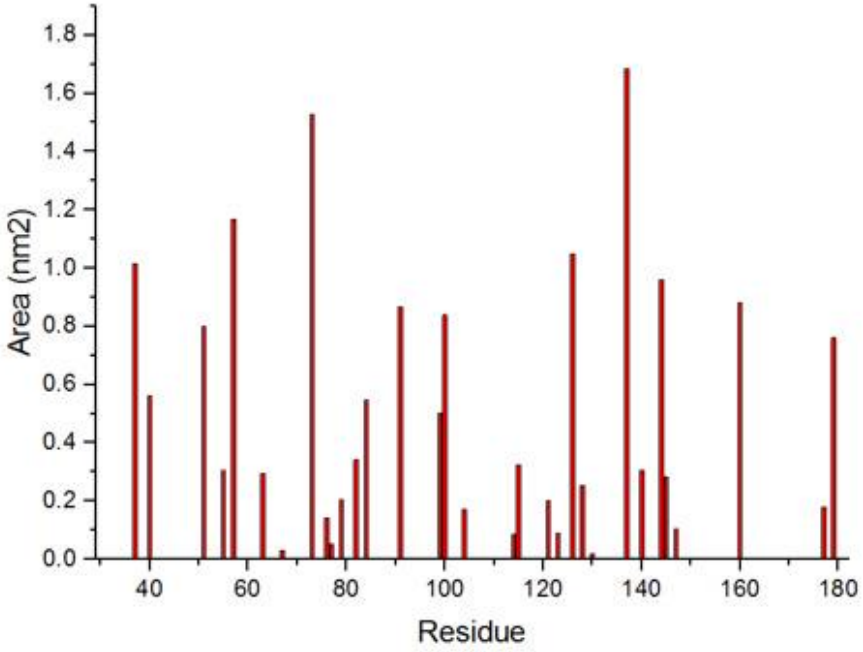

Figure 4.10. Calculated Solvent Assessible Surface Area (SASA) values per residue of L-PGDS in its apo form. SASA was calculated using gromacs tools over the average of three 200ns simulation repeats using values from the last 150ns of each simulation repeat. For clarity, only residues that were experimentally identified to have undergone perturbation in chemical shifts through NMR are shown.

In addition to NMR spectroscopy, our collaborators also performed Small Angle X-ray scattering in solution (SAXS) experiments for information about the L-PGDS- Aβ40 complex. SAXS is able to provide low resolution information about the system, describing details about overall shape of the complex. One quantity in which SAXS can provide is the radius of gyration (Rg). Our experimental collaborators found an approximately 0.1nm increase in the Rg of the L-PGDS- Aβ40 complex, compared to L-PGDS in its Apo form. We thus calculated the Rg values of L-PGDS- Aβ40 complex in L-PGDS in its apo form from MD simulations. We compared the Rg distributions between apo L-PGDS and our L-PGDS- Aβ40 complex model from BM1, and showed a similar corresponding shift in the distribution of the Rg values (Fig. 4.11). The Rg distributions were calculated over three 200ns classical MD simulation repeats using values from the last 150ns of each simulation repeat. The generation a theoretical scattering profile and comparing it with the SAXS experimental scattering profile[150] would be especially useful when large domain changes are observed upon binding between biomolecules. However, the Rg changes when Aβ40 binds to L-PGDS were considered to be small. Moreover, in this study, we are more focused on the fine atomistic interactions between L-PGDS and Aβ40. The SAXS data was used in this study to provide confidence for our model in that L-PGDS binds to Aβ40 in a 1:1 ratio, with no aggregation observed. Also, the SAXS space filling model generated of the L-PGDS-Aβ40 complex shows a globular domain with an additional domain near the mouth of the L-PGDS cavity that was not present in the absence of Aβ40. This serves to validate our model of L-PGDS directly binding to Aβ40 at the mouth of the L-PGDS cavity (Fig. 4.9).

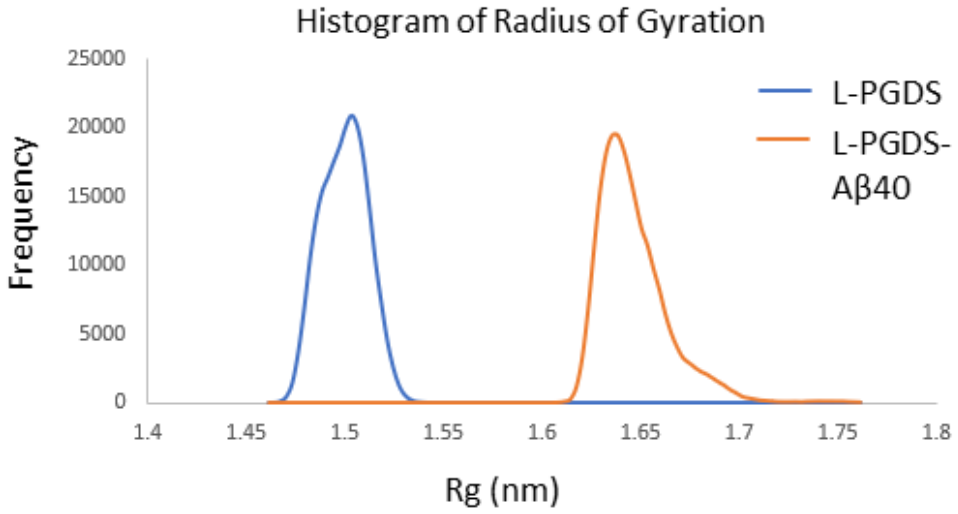

Figure 4.11 Histogram of Rg distribution between L-PGDS apo (blue) and L-PGDS- Aβ40 complex (orange). Rg values were calculated from three 200ns classical molecular dynamics simulations and using values from the last 150ns of each simulation repeat. The Rg values of the L-PGDS- Aβ40 complex were calculated from the classical molecular dynamics simulations initiated from binding mode 1.

4.3.7 Comparison of L-PGDS- Aβ complex and Anticalin US7- Aβ complex

Representative structures of Aβ40 in complex with L-PGDS corresponding to the three identified binding modes were obtained by simulation. The results suggest that the binding mode 1 (BM1) is the most favourable one. While L-PGDS and Anticalin US7 both contain the lipocalin fold, we do not expect their binding modes towards Aβ to be exactly the same. We sought to compare our model of L-PGDS in complex with Aβ40 with the crystal structure of Anticalin US7 in complex with Aβ (PDB ID: 4MVI). To do so, we superimposed our model of L-PGDS in complex with Aβ40 (BM1) onto the crystal structure of Anticalin US7 in complex with Aβ. It is observed that both L-PGDS and Anticalin US7 bind Aβ at a similar region (Fig. 4.12). However, there are differences observed in the interactions (Fig. 4.12), which could be attributed to the differences in amino acid sequences between L-PGDS and

Anticalin US7. These differences arise despite the initial coordinates of Aβ for the simulation of the L-PGDS- Aβ complex being derived from the coordinates of Anticalin US7 in complex with Aβ (PDB ID: 4MVI). In this aspect, MD simulation can be considered a tool for binding mode refinement of L-PGDS in complex with Aβ40.

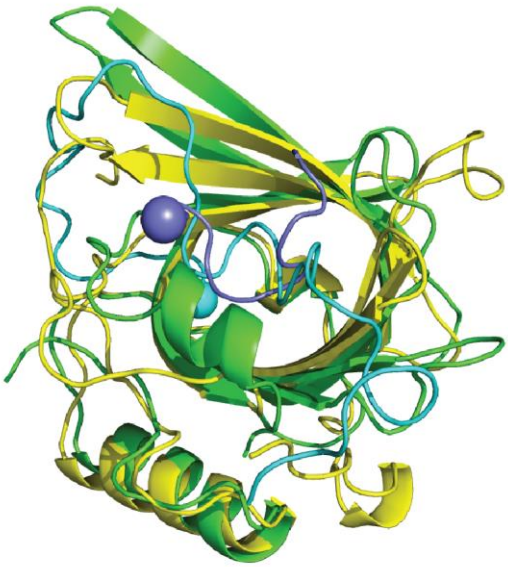

Figure 4.12. Superimposition of L-PGDS (Green Cartoon) -Aβ40 Binding Mode 1 model onto the crystal structure (PDB ID: 4MVI) of Anticalin US7 (Yellow Cartoon) in complex with Aβ (Purple Cartoon). For clarity, Aβ40 from Binding Mode 1 model is shown as cyan cartoon. N-Terminus of Aβ peptide shown as spheres.

4.3.8 Perspective of simulation findings

In many neurodegenerative diseases such as Alzheimer's Disease, Parkinson's disease, Huntington Disease etc., the abnormal deposition of misfolded peptides is a key pathological feature[13]. Therapeutics development targeting amyloids have been challenging and unsuccessful thus far[151, 152]. One possible reason for the failure of Aβ40 therapeutics is the lack of understanding of the biomolecular interactions between Aβ binders and Aβ40. This in-silico study provides the first model of L-PGDS interactions with Aβ40. We proposed several

biomolecular interactions that can be verified experimentally by our experimental collaborators. Studying the biomolecular recognition between L-PGDS and Aβ40 could aid in the development of therapeutics that prevents the aggregation of Aβ40 monomers into oligomers and fibrils, thus serving potentially as an early intervention or preventive therapeutic strategy.

In addition, a model of the L-PGDS- Aβ40 interactions may also have value to contribute to our mechanistic understanding of L-PGDS binding to its substrates. In a recent proteomic study, L-PGDS was able to extract and bind small molecule drugs[147]. This finding has important biological implications, because it suggests the potential of such compounds to compromise the beta chaperone function of L-PGDS. As an illustration, a preliminary study performed by our collaborators demonstrated that anti-cholinergic drugs may compromise the Aβ chaperone function of L-PGDS. Through NMR studies and docking calculations, binding models of the small molecule drugs to L-PGDS were generated. Superimposition of these binding models onto our model of the L-PGDS-Aβ40 complex showed a clash in the binding site of the drugs and Aβ40 binding site on L-PGDS (Data not shown).

4.4 Conclusion and future direction

In this study, we attempted to understand the atomistic interactions between L-PGDS and Aβ40. We used H-REMD enhanced sampling and a clustering method based on the identification of residue pairs between L-PGDS and Aβ40 with high contact probabilities to identify binding modes between L-PGDS and Aβ40. We have also subsequently showed the agreement of our simulation data with experimental NMR and SAXS data.

A future direction for this project would be to study the allosteric changes that arises from the binding of substrates to L-PGDS. Allosteric effects refer to changes that arise in a biomolecule

at regions away from the binding site[153]. This phenomenon is commonly observed in biology and involved in the regulation of many biological processes[154]. The possibility of allosteric effects in L-PGDS is proposed based on the presence of extensive chemical shift perturbations observed on L-PGDS residues when Aβ40 binds to L-PGDS. Another future direction of this project would be to understand the how the binding of small molecule drugs towards L-PGDS may alter or disrupt the Aβ chaperone function of L-PGDS.

5. The plasticity of *Arabidopsis thaliana* Profilin3 N-Terminal extension and its role in Formin poly-proline binding

5.1 Introduction

5.1.1 Profilins are proteins involved in plant actin cytoskeleton regulation

The formation of actin cytoskeleton is a key process in the regulation of cellular activities of plants, such as growth[155] and defence mechanisms[156]. To initiate the process of actin filament polymerization from actin monomers, a nucleation complex is formed consisting of actin monomers and an array of actin binding proteins[157]. Profilins are one of the mediators in plant actin assembly at the early stage, through their interactions with actin monomers[158]. Profilins (PRFs) direct bound actin monomers to the elongating termini of the filament, conferring directional polarity to actin polymerization[159, 160]. In addition to actin binding, they also are able to bind to regions of proteins containing proline repeats, known as poly-prolines [161]. One such binding partner of Profilins are known as formins[162]. Formin proteins are characterized by the presence of a formin homology 1 (FH1) domain, which contain proline rich regions [163]. This process, known as formin mediated actin assembly[164] is one of the key pathways in actin filament formation. Thus, the assembly of actin filaments is regulated through the dynamic associations and interactions of Profilins with formins and Actin.

5.1.2 Arabidopsis thaliana Profilin 3 is a unique profilin isoform containing N-Terminus extension

In Arabidopsis thaliana (At), there are 5 Profilin (PRFs) isoforms, AtPRF1-AtPRF5 [165]. Our experimental collaborators recently discovered through sequence alignment of the 5 AtPRFs that AtPRF3 is unique in that it contains a 37 residue intrinsically disordered N-Terminus extension (NTE) not present in other AtPRFs isoforms [166] (Fig. 5.1). AtPRF3 was found to bind with a stronger binding affinity towards At formin FH1 poly-proline segment

(RVPPPPPPPPPLP) (PolyP) as compared to other AtPRFs by microscale thermophoresis assay. Microscale Thermophoresis assay also showed that the binding affinity of AtPRF3 towards PolyP was dependent on the length of the NTE. Several constructs of the NTE were generated, namely AtPRF3-FL, consisting of the full length AtPRF3 with the entire NTE of 37 residues; AtPRF3Δ22 with NTE truncation at the 22nd residue (residues 23-37 of NTE present); and AtPRF3Δ37, without the 37 residues NTE. The Kd values showed a decreasing binding affinity in the order AtPRF3-FL, AtPRF3Δ22, and AtPRF3Δ37, towards AtFH1 PolyP. Biologically, AtPRF3 is shown to have a negative regulatory effect on AtFH1 mediated actin polymerization. Upon exposure Pathogen Associated Molecular Patterns, the levels of AtPRF3 in *Arabidopsis thaliana* are dynamically regulated as a mechanism to control formin mediated actin assembly[166]. To understand the atomistic basis for the role of the NTE in enhanced binding towards PolyP, it is important to study the biomolecular interactions between AtPRF3 and PolyP.

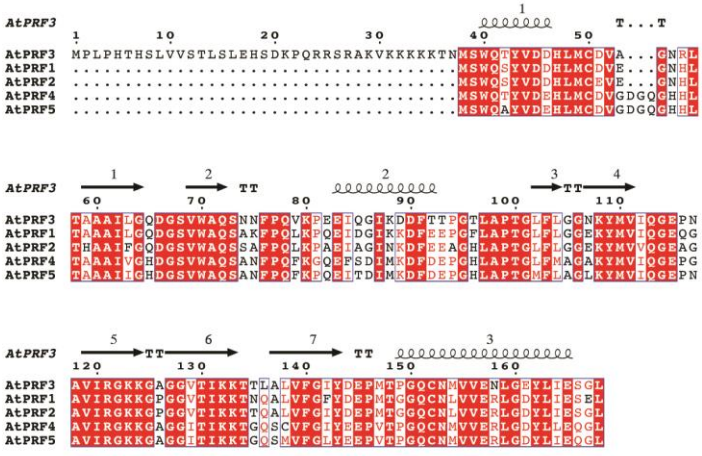

Figure 5.1. Sequence Alignment of the 5 AtPRFs isoforms. Alignment performed by clustal Omega. Identical residues are highlighted in red.

5.1.3 Experimental studies of interactions between Profilins and PolyP

Previous structural biology studies have studied the interaction between different profilins isoforms of various species and poly-proline ligands. Examples of these include the structure of human platelet profilin in complex with a poly-proline decamer [167] and structures of mouse Profilin2a in complex with poly-proline peptides from the vasodilator-stimulated phosphoprotein (VASP) and mammalian homologue 1 of Drosophila diaphanous (mDia1) proteins[168]. The profilin fold contains N- and C- terminal alpha helices with a beta sheet domain in between. The two Alpha helices form the Poly-proline binding site. The poly-proline adopts a type II polyproline helix[168]. However, these profilins lack the NTE found in AtPRF3, which is the interest of our present work.

To understand the interactions between AtPRF3 and PolyP and to study the role of NTE in enhancing PolyP binding, our experimental collaborators attempted to use X-Ray crystallography to solve the crystal structures of AtPRF3 and AtPRF2 in both the apo state and PolyP bound state. However, experimental determination of structures from X-Ray crystallography of structures is challenging, in which biomolecules have to be packed in a crystal lattice in a highly ordered state[43]. Because of the highly flexible nature of the AtPRF3 NTE, several truncated variants of the AtPRF3 NTE were generated to potentially stabilize the protein and facilitate the crystallization process. They attempted to crystalize AtPRF3-FL, consisting of the full length AtPRF3 with the entire NTE of 37 residues, AtPRF3Δ22 with NTE truncation at the 22nd residue (residues 23-37 of NTE present), and AtPRF3Δ37, without the 37 residues NTE. After multiple attempts, crystals of AtPRF3-FL were unable to be obtained, possibly because of the intrinsically disordered nature of the NTE. The structure of AtPRF3Δ37 without the NTE 37 residues, was solved (Fig. 5.2A). For the structure of AtPRF3Δ22, surface mutations (SM) of K33A/K34A/K35A on the NTE and an additional mutation of L136 to Q

had to be performed in order to obtain in order to obtain diffracting crystals (AtPRF3Δ22SMplusL136Q). However, residues 23 to 30 (QRRSRAKV) on the NTE of this structure could not be modelled as the electron density was unclear for that region. For clarity, we designate the naming of this crystal structure as AtPRF3Δ30SMplusL136Q (Fig. 5.2B). In this structure, the NTE residues from K31 to N37 were observed in an extended conformation (Fig. 5.2B, magenta). The crystal structure of AtPRF2 in complex with PolyP was also solved (Fig. 5.3C). It is observed that the core region of AtPRF3Δ37 (Fig 5.2A), AtPRF3Δ30SMplusL136Q (Fig. 5.2B), and AtPRF2 (Fig. 5.2C) are conserved, consisting of seven anti-parallel beta stranded sheets sandwiched by four α-helices (X-Ray Crystallography performed by Mr Qiao Zhu, A/Prof Gao Yonggui's lab). Since the structure of AtPRF3 in complex with PolyP could not be solved, we attempted to visualize the interactions of AtPRF3 and PolyP. The crystal structure of AtPRF2 in complex with PolyP was superimposed onto AtPRF3Δ30SMplusL136Q (Fig. 5.2D). The AtPRF2-PolyP complex is presumed to be a representative conformation of the interaction between AtPRFs and PolyP, since the core region structures of AtPRF3 and AtPRF2 are conserved (Fig. 5.2B, 5.2C). It was observed that the NTE of AtPRF3Δ30SMplusL136Q occupied the PolyP binding site, forming a structural clash (Fig. 5.2D).

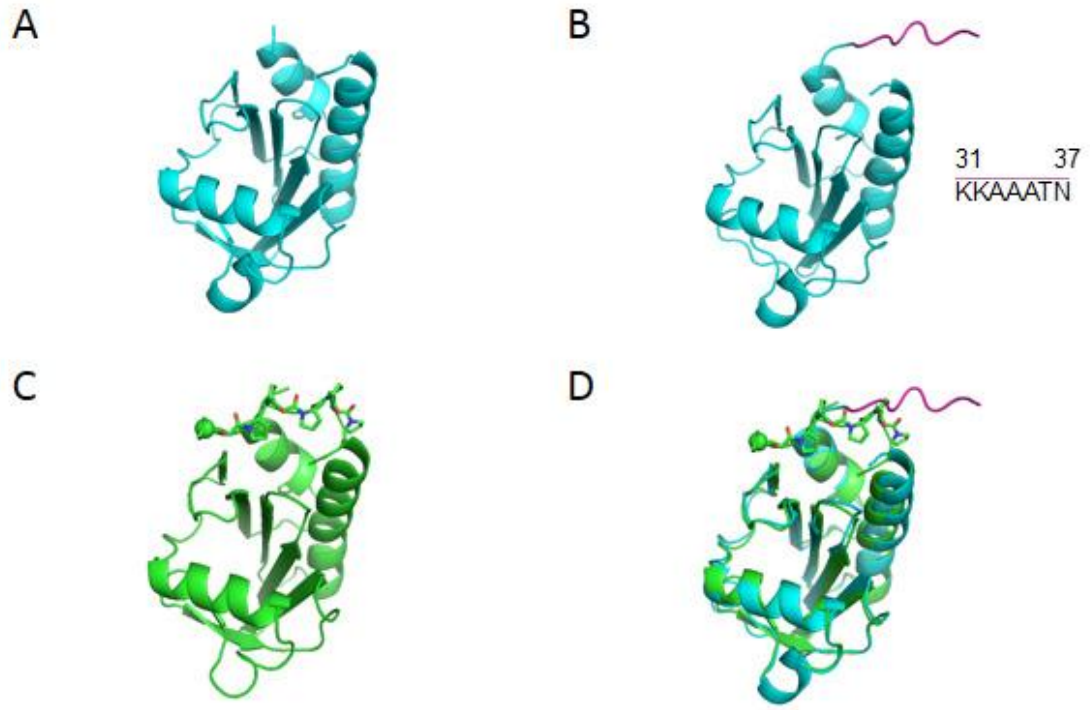

Figure 5.2. Crystal structures of AtPRFs. A. AtPRF3Δ37. B. AtPRF3Δ30SMplusL136Q. C. AtPRF2-PolyP complex. D. Superimposition of AtPRF2 in complex with PolyP (C) onto AtPRF3Δ30SMplusL136Q (B) reveals structure clash between N-Terminal Extension (NTE) of AtPRF3Δ30SMplusL136Q (B) (magenta cartoon) and PolyP binding site of AtPRF2. AtPRF3 shown as cyan cartoon, AtPRF3Δ30SMplusL136Q NTE shown as magenta cartoon, AtPRF2 shown as green cartoon, AtPRF2-PolyP shown as green sticks. N-terminal of PolyP shown as green spheres.

Our biochemistry collaborators validated the function of the crystallization generated construct, AtPRF3Δ22SMPlusL136Q, through a series of experiments. Microscale Thermophoresis assay to measure the dissociation constant (Kd) of AtPRF3Δ22, AtPRF3Δ22SM, and AtPRF3Δ22SMPlusL136Q to PolyP showed the presence of the L136Q mutation negligibly affects binding to PolyP. AtPRF3Δ22, AtPRF3Δ22SM, and AtPRF3Δ22SMPlusL136Q variants were able to inhibit formin mediated actin assembly in similar manners. In vivo time lapse imaging to monitor the functions of different AtPRF3Δ22 NTE variants in yeast showed no significant differences in actin cable elongation speed (Biochemistry experiments performed by Mr. Sun He, Asst Prof Miao Yansong's lab). AtPRF3Δ22 is shown to be able to

maintain in-vitro and in-vivo function, while having a moderate decrease binding affinity compared to full length AtPRF3[169].

5.1.4 Aim of Project and scope

The aim of this project is to understand the biomolecular recognition between AtPRF3 and poly-proline (PolyP) segment (RVPPPPPPPPPLP) from the *Arabidopsis thaliana* FH1. Since the crystal structure of AtPRF3 in complex with PolyP could not be obtained experimentally, molecular modelling through molecular dynamics simulations can be applied for the studying of biomolecular recognition. Understanding the biomolecular interactions between AtPRF3 and PolyP provides insights into the dynamic regulation of formin mediated actin assembly to regulate processes in plant defence and developmental mechanisms.

When we superimposed AtPRF3Δ30SMplusL136Q onto the AtPRF2-PolyP complex, it was observed that residues K31-N37 belonging to the NTE of AtPRF3Δ30SMplusL136Q occupied the PolyP binding pocket, forming a structural clash with bound PolyP. Such a comparison could be made because the profilin core of AtPRF2 and AtPRF3 is highly conserved. To resolve this apparent discrepancy, we hypothesized that given the intrinsic disordered and flexible nature of the AtPRF3 NTE, the NTE is able to adopt various dynamic conformations that can be rearranged to allow the binding of PolyP. We used Hamiltonian Replica-Exchange Molecular Dynamics (H-REMD) simulations[64] to sample the conformation space of AtPRF3 NTE. This was performed using the crystal structure of AtPRF3Δ30SMplusL136Q, with residues K31 to N37 resolved, and L136Q mutation corrected to L136 during modelling. We designate this model structure as AtPRF3Δ30SM. Our results reveal the conformational plasticity of the NTE, which adopts a “closed” conformation that would be able to adopt “open” conformations to enable the binding of PolyP.

In the second part of this project, we sought to understand the biomolecular interactions between AtPRF3 and AtFH1 PolyP (RVPPPPPPPPPLP). To do so, we modelled the remaining unresolved residues of NTE and used classical molecular dynamics simulations to sample the conformation space of AtPRF3 in complex with PolyP. Following which, we performed geometric clustering and binding free energy calculations to identify representative conformations of the AtPRF3-PolyP complex. We focused our studies using AtPRF3Δ22SM truncation instead of AtPRF3-FL. This is because simulating the full length of the NTE requires higher computational complexity, which is prohibitive given our availability of computational resources, and compounded by the intrinsically disordered nature of the NTE. Moreover, it has been experimentally demonstrated that AtPRF3Δ22 was able to maintain biological function in-vitro and in-vivo (Section 5.1.3 Experimental studies of interactions between Profilins and PolyP)[169]. Additionally, to also investigate the effects of charged residues on the NTE, we performed simulations on another AtPRF3Δ22 NTE variant where charged residues are mutated to alanine residues (AtPRF3Δ22SM3RtoA). Our simulation findings show that the NTE binds to PolyP adaptively involving multiple binding modes, maintaining binding affinities towards PolyP. Our simulation findings are supported by experimental results, and consistent with the conformationally flexible nature of intrinsically disordered regions. We also constructed and evaluated an alternate model of the AtPRF3Δ22SM-PolyP complex where the NTE was modelled with partial helical secondary structure.

For clarity, we have summarized the various systems discussed in this project in the form of a Table (Table 5.1). We define each system, the presence or absence of PolyP ligand, whether each conformation is a crystal structure or model, the type of simulations performed, as well as provide any descriptive remarks.

Table 5.1 Summary of various AtPRFs systems discussed in this chapter

Name	NTE	Ligand	Crystal/ Model	Simulation	Remarks
AtPRF2-PolyP	-	PolyP	Crystal	-	-
AtPRF3FL	Full Length	-	-	-	Could not be crystallized
AtPRF3Δ37	-	-	Crystal	-	-
AtPRF3Δ37-PolyP	-	PolyP	Model	cMD	AtPRF3Δ37-PolyP complex obtained by superimposing AtPRF2-PolyP complex onto AtPRF3Δ37
AtPRF3Δ22SMplusL136Q	Q23-N37, (K33A/K34A/K35A)	-	-	-	Crystallization construct
AtPRF3Δ30SMplusL136Q	K31-N37, (K33A/K34A/K35A)	-	Crystal	-	Crystal from the crystallization construct of AtPRF3Δ22SMplusL136Q, but only residues K31 to N37 of NTE resolved
AtPRF3Δ30SM	K31-N37, (K33A/K34A/K35A)	-	Model	H-REMD	Model generated by converting L136Q of AtPRF3Δ30SMplusL136Q to L136
AtPRF3Δ22SM-PolyP	Q23-N37, (K33A/K34A/K35A)	PolyP	Model	cMD	Remaining residues of NTE modelled, PolyP conformation from superimposition to AtPRF2-PolyP structure
AtPRF3Δ30SM3RtoA-PolyP	Q23-N37, (K33A/K34A/K35A), (R24A/ R25A/ R27A)	PolyP	Model	cMD	Similar to AtPRF3Δ22SM-PolyP, 3 charged Arginines on NTE converted to Alanines
AtPRF3Δ22-PolyP	Q23-N37 (K33, K34, K35)	PolyP	Model	cMD	Similar to AtPRF3Δ22SM-PolyP, 3 Alanines of the SM NTE converted to Lysines. The "Wild Type" AtPRF3Δ22 NTE

5.2 Method

5.2.1 Hamiltonian Replica Exchange Molecular Dynamics to sample to conformation of AtPRF3 NTE

To generate the initial model of AtPRF3Δ30SM for Hamiltonian Replica Exchange Molecular Dynamics (H-REMD), the L136Q mutation was corrected to L136 for Chain A in the structure AtPRF3Δ30SMplusL136Q. AtPRF3Δ30SM contained residues K31 to N37 of the NTE

resolved. Replica Exchange with Solute Scaling (REST2)[64], a variant of H-REMD simulation was performed using Gromacs[58] 5.1.4 with Plumed[114] 2.4 plugin. The residues K31-N37 (KKAAATN) belonging to the NTE were selected for Hamiltonian Scaling. The Charmm36M[56] force field and Charmm modified TIP3P[105] water model was used for the simulation. The protein was solvated in a cubic box with a distance of 1.2nm from the box edge. Counter-ions were added to neutralize the system. The LINCS[111] algorithm was used to constrain bonds containing hydrogen atoms, enabling a time step of 2fs. Particle Mesh Ewald[112] was implemented. A cutoff of 1.2nm for electrostatics and van der waal's interactions were used. The temperature of the system was maintained using the V-rescale[113] thermostat. Steepest Descent energy minimization was performed for 1000 steps to remove any initial bad contacts. A 1ns equilibration in the NVT ensemble was performed prior to starting the simulation. Ten replicas were used for the simulation and an effective temperature range of 300K-600K was used to generate the Hamiltonian scaling factors following a geometric progression. The temperatures are 300, 324.018, 349.959, 377.976, 408.237, 440.92, 476.22, 514.346, 555.525, and 600K. This resulted in an average exchange probability of approximately 30%. Exchanges were attempted every 2ps, and the simulation was performed for 800ns. Coordinates were saved every 2ps.

5.2.2 Free Energy Surface to describe conformational Plasticity of AtPRF3 NTE

Analysis of the simulation was performed on the replica with the unscaled potential energy (with effective temperature=300K). The first 100ns of the trajectory was discarded to take into account the effects of initial equilibration. In order to understand the conformational dynamics of the AtPRF3-NTE, we plotted the free energy surface as a function of the dihedral angle formed by the alpha carbons of AtPRF3 Residues 36-39 against the center of the mass distance between the NTE and the C-terminal Helix (CTH). These reaction coordinates characterized

the dynamics of the NTE and were chosen after extensive detailed observation and analysis of the H-REMD simulation trajectory. The dihedral angle formed by the alpha carbons of residues 36-39 characterizes the "hinge" motion of the NTE, while the center of the mass distance between the NTE and the CTH is an indicator if the NTE would occlude the PolyP binding site. The dihedral angle calculations were done using the gmx_angle tool, while the center of mass distance calculations were performed with the gmx_distance tool, both inbuilt into the Gromacs[58] package. In order to generate the free energy surface, the projection was divided into bins, and the counts in each bin were used to calculate its relative free energy by the formula $E = -RTln(P_i/P_0)$. In this formula, P_i refers to the number of counts in each bin, P_0 refers to the count of the most populated bin, R refers to the gas constant and T refers to the temperature in K (300K). To evaluate the convergence of the simulations, we plotted the free energy surface with different time blocks throughout the simulation and evaluated the conservation of the Free Energy Surface.

To characterize each minima, we extracted their respective conformations belonging to their respective regions. Geometric clustering was used to identify their representative conformation. Clustering of the NTE was performed using the gmx_cluster utility of the Gromacs package, using the GROMOS algorithm with a cutoff of 0.5nm. To further visualize each representative conformation, we superimposed the representative structure of each minima onto the experimentally resolved crystal structure of AtPRF2 (ChainA) containing the PolyP segment (Chain C) to evaluate the possibility of a structural clash between the AtPRF3-NTE and the PolyP segment, as well as an occlusion of the PolyP binding site by the AtPRF3-NTE. The AtPRF3-NTE conformation is considered to be in an "open" state if it does not structurally clash with the PolyP structure from AtPRF2.

5.2.3 Model construction of AtPRF3Δ22- PolyP complex

To construct a model of AtPRF3Δ22SM (containing NTE residues Q23 to N37) in complex with PolyP, the remaining unresolved residues from QRRSRAKV (Q23-V30) were modelled in an extended and disordered conformation using Discovery Studio[107]. The structure of AtPRF3Δ30SM had residues KKAAATN (K31-N37) of the NTE resolved. The starting conformation of AtPRF3Δ30SM used to extend the NTE was the representative structure with an "open" conformation of the AtPRF3Δ30SM's NTE (K31-N37) from the Free Energy Surface. The modelled AtPRF3Δ22SM structure was superimposed onto the experimental crystal structure of AtPRF2 (chain A) in complex with PolyP (chain C). The coordinates of the PolyP segment together with aligned AtPRF3Δ22SM was extracted in order to build the complex model of AtPRF3Δ22SM with the PolyP segment. In the experimental structure of PolyP, the N terminal Arginine residue was not resolved and was hence modelled into the structure using discovery studio 4.1[107].

In order to investigate the contribution of charged residues in the NTE towards PolyP binding, we performed additional simulations of AtPRF3 with different NTE variants in complex with PolyP. In addition to the model of AtPRF3Δ22SM, we generated the additional AtPRF3Δ22 and AtPRF3Δ22SM3RtoA variant. To generate AtPRF3Δ22SM3RtoA model, three arginine residues R24, R25 and R27 of AtPRF3Δ22SM were converted to alanine residues. To generate the AtPRF3Δ22 wild type variant, the AtPRF3Δ22SM model had residues A33, A34 and A35 converted to K. Prior to simulations setup, the additional AtPRF3Δ22 variant models were subjected to energy minimization in-vacuo using a similar energy minimization protocol as above H-REMD simulations (Section 5.2.1 Hamiltonian Replica Exchange Molecular Dynamics to sample to conformation of AtPRF3 NTE). The crystal structure of AtPRF2-PolyP

complex was superimposed onto the models AtPRF3Δ22 NTE variants in order to extract coordinates of the AtPRF3Δ22 NTE variants in complex with PolyP.

5.2.4 Classical Molecular Dynamics simulations and binding mode analysis of PolyP bound AtPRF3Δ22

Classical molecular dynamics simulations were initiated, with topology generation, energy minimization and equilibration performed with a protocol similar to as described above for the H-REMD simulations (Section 5.2.1 Hamiltonian Replica Exchange Molecular Dynamics to sample to conformation of AtPRF3 NTE). To sample the conformational space of model AtPRF3Δ22-PolyP complexes, we performed three repeats of 100ns production simulations with different initial velocities at 300K. The last 50ns of each simulation was extracted and combined for binding mode analysis to allow a larger sampling of the conformational space.

We performed geometric clustering by the full protein backbone of the combined trajectory using the Gromos algorithm with the in-built gromacs gmx cluster tool, using a cutoff of 0.25nm. In this way, similar conformations are grouped into clusters with distinct conformations that we can further analyse. The analysis was performed for clusters contributing to more than 15% of the population. The representative structures for each cluster were further analysed to study the binding between AtPRF3Δ22 and PolyP in order to understand the contribution of the NTE to PolyP binding. Each cluster is assigned a label based on the NTE variant and based on cluster population size (Eg. AtPRF3Δ22SM-cluster 1).

To understand the interactions between the AtPRF3 NTE and PolyP, biomolecular interactions were identified based on definitions of various biomolecular interactions (Section 2.6. Identification of Biomolecular Interactions). Contact occupancy of each biomolecular

interaction was calculated for each cluster of the various binding modes. The occupancies were calculated as a percentage of the total number of frames belonging to that particular cluster. Contacts with occupancies higher than 25% were included.

5.2.5 Binding Energy analysis

Using the Molecular Mechanics-Poisson Boltzmann Surface Area (MM-PBSA) method, the binding free energy is defined as $E_{Binding} = E_{vdw} + E_{elec} + E_{Polar} + E_{Apolar} - T\Delta S$, whereby E_{vdw} refers to van der Waal's interaction energy, E_{elec} refers to electrostatic interaction energy, E_{Polar} refers to polar solvation energy approximated by the Poisson Boltzmann equation, E_{Apolar} refers to the Apolar solvation energy approximated using a Solvent Accessible Surface Area (SASA) model, while ΔS refers to solute entropy. A common approach involves calculating binding energy without considering entropy. The polyproline helix is known to be rigid and we assume similar entropy contributions to allow meaningful comparisons of binding free energies. Moreover, the explicit calculation of entropy is inaccurate and computationally expensive[71]. G_mmpbsa tool was used to calculate binding energy calculations and bootstrapping analysis with 2000 steps were used to calculate errors in free energy calculations. We also set up a parallel simulation calculation of AtPRF3Δ37-PolyP system without the NTE following a similar protocol as that performed for the other AtPRF3Δ22 NTE. By comparing the binding energies of AtPRF3Δ22 variants containing the NTE and AtPRF3Δ37 without the NTE towards PolyP, we can understand the contribution of the NTE to the enhanced binding of AtPRF3 towards PolyP.

5.3 Results and Discussion

5.3.1 Plasticity of the AtPRF3 NTE revealed by H-REMD

When we superimposed the AtPRF2-PolyP complex onto the AtPRF3Δ30SMplusL136Q structure, the Profilin core was found to be highly conserved. It is observed that residues K31 to N37 of the NTE in AtPRF3Δ30SMplusL136Q occupied the PolyP binding pocket. It is hypothesized that AtPRF3-NTE is able to rearrange itself to form an open cavity, thereby allowing the binding of PolyP. From AtPRF3Δ30SMPlusL136Q structure, we converted the L136Q mutation to L136 and designate the model as AtPRF3Δ30SM since residues K31-N37 of the NTE were resolved. By performing H-REMD simulation of AtPRF3Δ30SM, we examined the conformational flexibility of the NTE.

The H-REMD simulation data is summarized in the free energy surface that was shown as a function of the dihedral angle formed by α carbons of the residues 36-39 against the center of the mass distance between the NTE and the C terminal helix (CTH). Seven local minima (Fig. 5.3A) and their respective representative structures were extracted by geometric clustering (Fig. 5.3B), describing the conformational states of the AtPRF3 NTE.

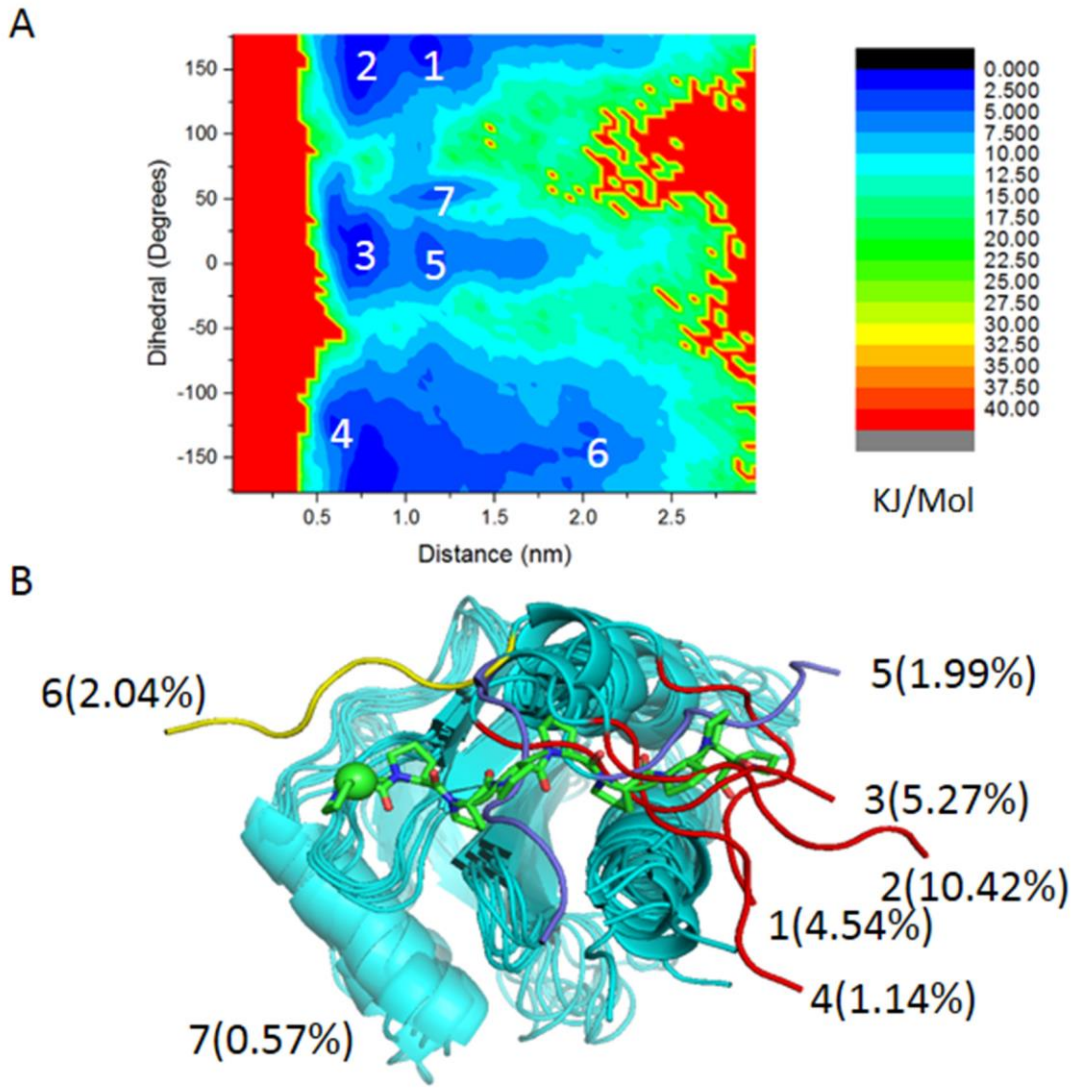

Figure 5.3. A. Free energy surface to characterize the dynamics of the N-Terminal Extension. It was plotted as a function of dihedral angle formed by C-Alpha carbons of residues 36 to 39 (Dihedral) against center of mass distance of the N-Terminal Extension and C-terminus helix (Distance). The free energy surface was generated through Hamiltonian Replica Exchange Molecular Dynamics simulation using the last 700ns of simulation data. Seven local minima of low relative free energies were identified and labelled 1-7. B. Representative structures of each local minima (cyan) was obtained through geometric clustering and superimposed onto AtPRF2-PolyP crystal structure (green). For clarity, only the PolyP segment of AtPRF2-PolyP crystal structure was shown (green sticks). The minima can be classified to be an "open" (yellow, minima6), "semi-open" (purple, minima 5 and 7), "closed" (red, minima 1-4) based on whether there is a clash of the NTE with the PolyP segment. Percentages refer to the ratio of the number of frames belonging to a particular minima to the total number of frames we sampled.

By assuming AtPRF3 would bind the PolyP in a similar way as AtPRF2, we have superimposed the structures of AtPRF3Δ30SM with the NTE onto that of the PolyP bound AtPRF2. In such

a way, the PolyP was “docked” onto the AtPRF3Δ30SM structure in the exact way as it binds to AtPRF2. Each representative structure of the local minima is defined as an “open” state of the AtPRF3 NTE if the structural clash is absent between the AtPRF3 NTE and the PolyP, otherwise, a “closed” state of the AtPRF3 NTE would be defined. We found that minimum 1,2,3 and 4 represent “closed” conformations of the AtPRF3Δ30SM NTE, whereby the NTE forms structural clashes with the PolyP segment after superimposition (Fig. 5.3B). Minimum 6 represents an “open” conformation of the AtPRF3Δ30SM NTE, whereby there is no structural clash between the AtPRF3Δ30SM NTE and the PolyP segment (Fig. 5.3B). In the representative structures of minima 5 and 7, the superimposed AtPRF3 NTE forms a minor structural clash with the PolyP segment of AtPRF2. Notably, the PolyP binding site is not occluded by the AtPRF3Δ30SM NTE, and the PolyP segment fits in the aligned PolyP binding site of AtPRF3. Given the dynamic nature of the AtPRF3Δ30 NTE, it is expected that upon PolyP binding in such conformations, the AtPRF3Δ30 NTE could alter its conformation to accommodate PolyP. We hence consider Minima 5 and 7 to be representing “semi-open” conformations of the AtPRF3Δ30SM NTE (Fig. 5.3B). We also used various simulated time windows from 100 to 550 ns, 650 ns, and 800 ns to show that the simulation of NTE is converged (Fig. 5.4).

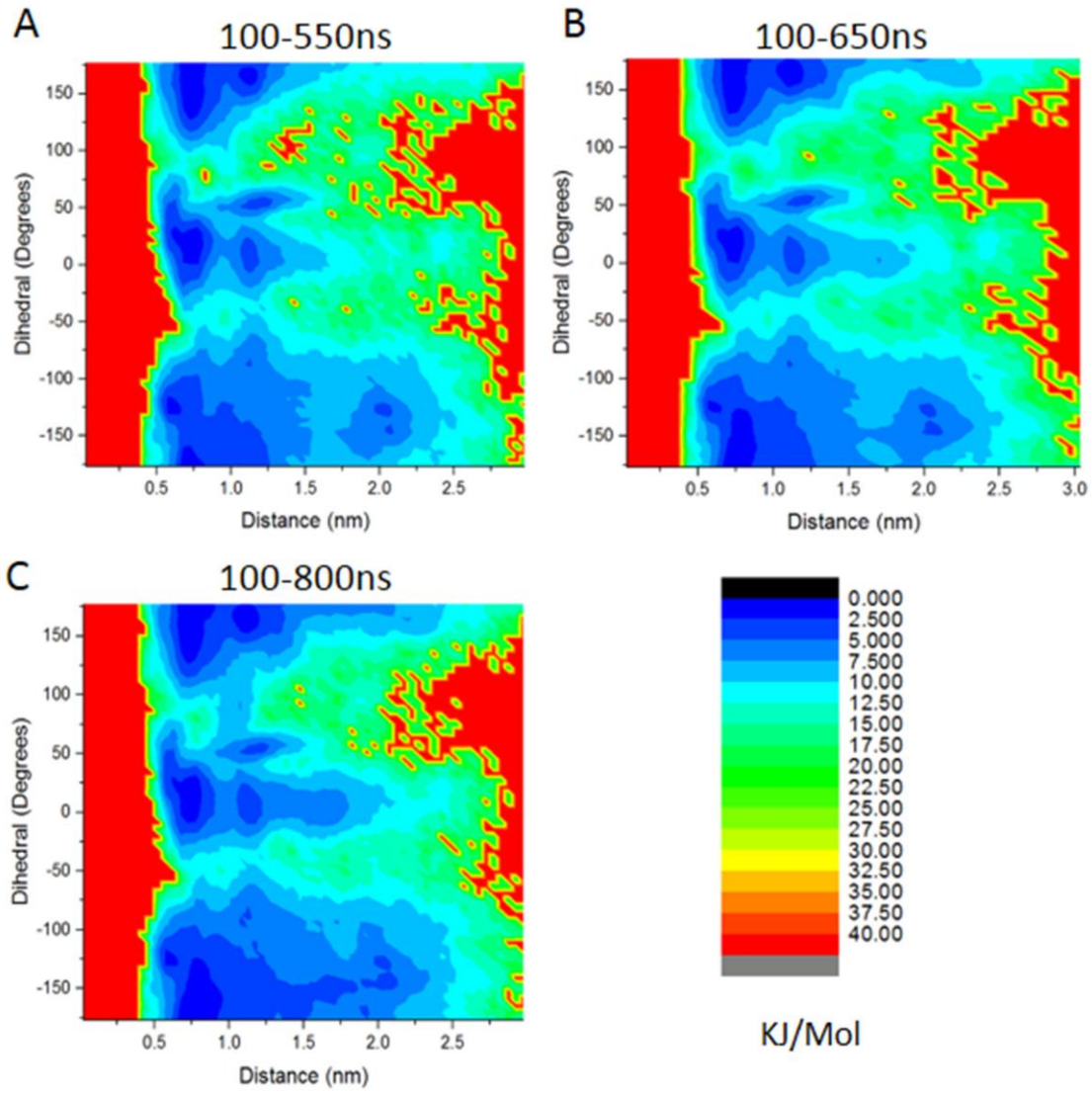

Figure 5.4. Convergence of H-REMD simulation by plotting various time blocks of the free energy surface, from A. 100ns to 550ns, B.100ns to 650ns, and C. 100ns to 800ns.

The free energy landscape reveals the plasticity and dynamic nature of the AtPRF3Δ30SM NTE in its apo form, which enables it to adopt diverse conformations, such as exposing or dynamically occluding the PolyP binding site. This suggests that the conformations of the AtPRF3Δ30SM NTE could be fine-tuned and regulated to allow the binding of PolyP during biological situations, such as plant immune responses in order to regulate actin filament assembly[169].

5.3.2 Understanding the interactions in the AtPRF3Δ37-PolyP complex

The interactions between AtPRF3Δ37 (without the NTE) and PolyP was studied previously by Sun et al using homology modelling and Molecular Dynamics simulations [166]. They used the apo structure of AtPRF1 (PDB ID 1A0K) as a homology modelling template and modelled AtFH1-PolyP (RVPPPPPPPPPLP) (PolyP) as a type 2 polyproline helix. Based on observation of the crystal structures of mouse Profilin2a in complex with poly-proline peptides from the vasodilator-stimulated phosphoprotein (VASP) and mammalian homologue 1 of Drosophila diaphanous (mDia1) proteins[168], they modelled the alignment of PolyP as either parallel (Termed "L" conformation for lower conformation) or perpendicular (Termed "U" conformation for upper conformation) to the N- and C- terminal helices axes. By the observation that Profilins can bind PolyP in forward and reverse orientations of the N-terminal[170], they modelled PolyP with forward (termed "NC" orientation) and reverse (termed "CN" orientation) alignments, relative to the PolyP N terminal position from the crystal structure of mouse Profilin2a in complex with poly-proline peptides from mammalian homologue 1 of Drosophila diaphanous (mDia1)[168]. They performed molecular dynamics simulations, geometric clustering and binding free energy calculations to evaluate different combinations of alignments and orientations of PolyP towards AtPRF3[166].

Since crystal structures of the AtPRF3Δ37-PolyP complex could not be resolved, we attempted to understand the interactions between AtPRF3Δ37 and PolyP. To do so, we superimposed the crystal structure of AtPRF3Δ37 onto the crystal structure of AtPRF2-PolyP complex. Our AtPRF3Δ37-PolyP model generated from crystal structure superimposition corresponds to the "U-NC" binding mode simulated by Sun et al[166]. Unlike the various suggested binding alignments[168] and orientations[170], we were able to only observe a single binding alignment and orientation from our crystal structures[169]. A key form of biomolecular

interactions between the profilin core and PolyP were found to be CH-π interactions[168]. CH-π interactions occur between the aromatic side chains of amino acid residues and the polarized C-H bonds of proline residues of PolyP. In our model of AtPRF3Δ37-PolyP complex generated by superimposing crystal structures of AtPRF3Δ37 and AtPRF2-PolyP, we were able to identify several CH-π interactions, in which aromatic residues Trp40, Tyr43, His47, Trp70, Tyr162 in close proximity of 0.5nm to Proline residues (Fig.5.5). Interestingly, the identified residues on the AtPRF3 core are found to be conserved throughout the 5AtPRFs (Fig. 5.1). Since our findings are consistent with the previous findings of Sun et al [166] and that the biomolecular interactions between AtPRF3Δ37 and PolyP have been extensively characterized in that work, we focused our subsequent work on the role of the AtPRF3 NTE.

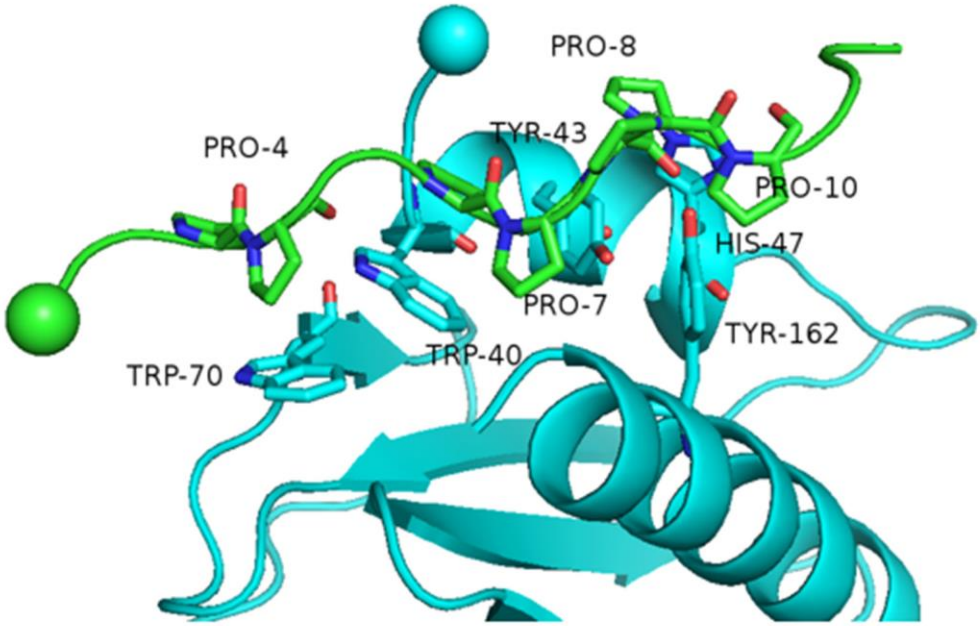

Figure 5.5. Binding pose of AtPRF3Δ37-PolyP complex. This was generated by superimposing AtPRF3Δ37 onto AtPRF2-PolyP complex. AtPRF3Δ37 shown as cyan cartoon. Aromatic residues in the profilin core shown as sticks and labelled. PolyP shown as green cartoon. Proline residues in close proximity of AtPRF3 aromatic residues are shown as green sticks, suggesting the identification of CH- π interactions.

5.3.3 Understanding the interactions between PRF3Δ22 and PolyP-The binding of AtPRF3 NTE to PolyP may involve multiple, adaptive binding modes

The previous biochemical data showed that AtPRF3 NTE enhanced the binding affinity towards PolyP. We were therefore motivated to dissect further the biomolecular interactions between the Q23-N37 of AtPRF3Δ22SM and PolyP, in which the structure of Q23-V30 was unresolved. To do so, we modelled the NTE residues in an extended and disordered conformation using Discovery studio 4.1[107]. We performed three repeats of 100ns classical molecular dynamics simulations with differing random initial velocities in order to enhance the sampling of the conformation space. Following which, we combined the last 50ns of each simulation repeat and performed geometric clustering to identify representative populations of clusters. Contact occupancies for identified biomolecular interactions were calculated. In addition, the binding free energies of AtPRF3Δ22SM towards PolyP was calculated. We generated the AtPRF3Δ22 variant by converting residues A33, A34 and A35 converted to K residues from the AtPRF3Δ22SM model. To investigate the effect of charged residues on the NTE towards the binding of PolyP, we generated the AtPRF3Δ22SM3RtoA variant, in which three arginine residues R24, R25 and R27 of AtPRF3Δ22SM were converted to alanine residues. Simulation and analysis was performed for these AtPRF3Δ22 variants in complex with PolyP following a similar protocol as that of the AtPRF3Δ22SM-PolyP complex.

The residues of the NTE was modelled in an extended and disordered conformation. Firstly, the crystal structure AtPRF3Δ30SM has the NTE residues from K31 to N37 in a disordered conformation. Also, Circular dichroism (CD) spectroscopy was performed by our collaborators to examine the secondary structures of AtPRF3Δ37 and AtPRF3 in apo and PolyP bound forms. In the results, AtPRF3Δ37 and AtPRF3 showed similar CD spectra profiles, either in the absence or in the presence of PolyP. It suggests that PolyP binding does not induce an apparent

change in secondary structures of AtPRF3 NTE. CD spectroscopy detected a slightly higher helix signal of PolyP bound AtPRF3 proteins than their apo forms, which can be possibly attributed to the detected helical conformation of PolyP[171] (Experiment performed by Mr. Qiao Zhu, A/Prof Gao Yonggui's lab)[169].

The representative clusters of the PolyP bound AtPRF3Δ22SM, AtPRF3Δ22SM3RtoA and AtPRF3Δ22 was obtained from molecular dynamics simulations and geometric clustering. One cluster of AtPRF3Δ22SM, one cluster of AtPRF3Δ22SM3RtoA, and two clusters of AtPRF3Δ22 conformations towards PolyP were identified. The representative conformations for each binding mode of AtPRF3Δ22 variants in complex with PolyP and disordered conformations of the NTE are shown in Figure 5.6. Identified biomolecular interactions between the NTE and their respective calculated contact occupancies are shown in Table 5.2. It is observed that the different variants of the NTE adopt different binding modes towards PolyP, involving different residues and biomolecular interactions (Fig. 5.6, Table 5.2). The conformations of each binding mode were extracted for binding free energy calculations using the MM-PBSA method[70] (Table 5.3). For comparison, simulations, clustering and binding energy analysis were also performed for the PolyP bound AtPRF3Δ37, in which the NTE is absent (Table 5.3). Calculated binding free energies of the AtPRF3Δ37-PolyP complex is -251.38KJ/Mol, consistent with the previous AtPRF3Δ37-PolyP "U-NC" calculation of Sun et al of -246.45KJ/Mol[166]. The binding energies to the PolyP bound AtPRF3Δ22SM, AtPRF3Δ22SM3RtoA, and AtPRF3Δ22 (Wild Type) clusters are all lower than those of AtPRF3Δ37 (Table. 5.3), which indicates a positive contribution of NTE towards PolyP binding, consistent with experimental findings[166]. Interestingly, the binding free energy calculations of AtPRF3Δ22SM, AtPRF3Δ22SM3RtoA and AtPRF3Δ22 clusters showed comparable binding free energies, that were all lower than AtPRF3Δ37, towards PolyP (Table

5.3). These findings are supported by microscale thermophoresis binding assay by our collaborators, whereby AtPRF3Δ22SM, AtPRF3Δ22SM3RtoA and AtPRF3Δ22 variants showed marginal differences in binding affinity towards PolyP (Experiments performed by Mr. Sun He, Asst Prof Miao Yan Song's lab). Taken together, the similar binding affinities but different binding modes and contact occupancies of AtPRF3Δ22 NTE variants towards PolyP suggests the NTE adopts non-specific and transient interactions, that may interconvert under different environmental conditions to maintain continuous binding towards PolyP.

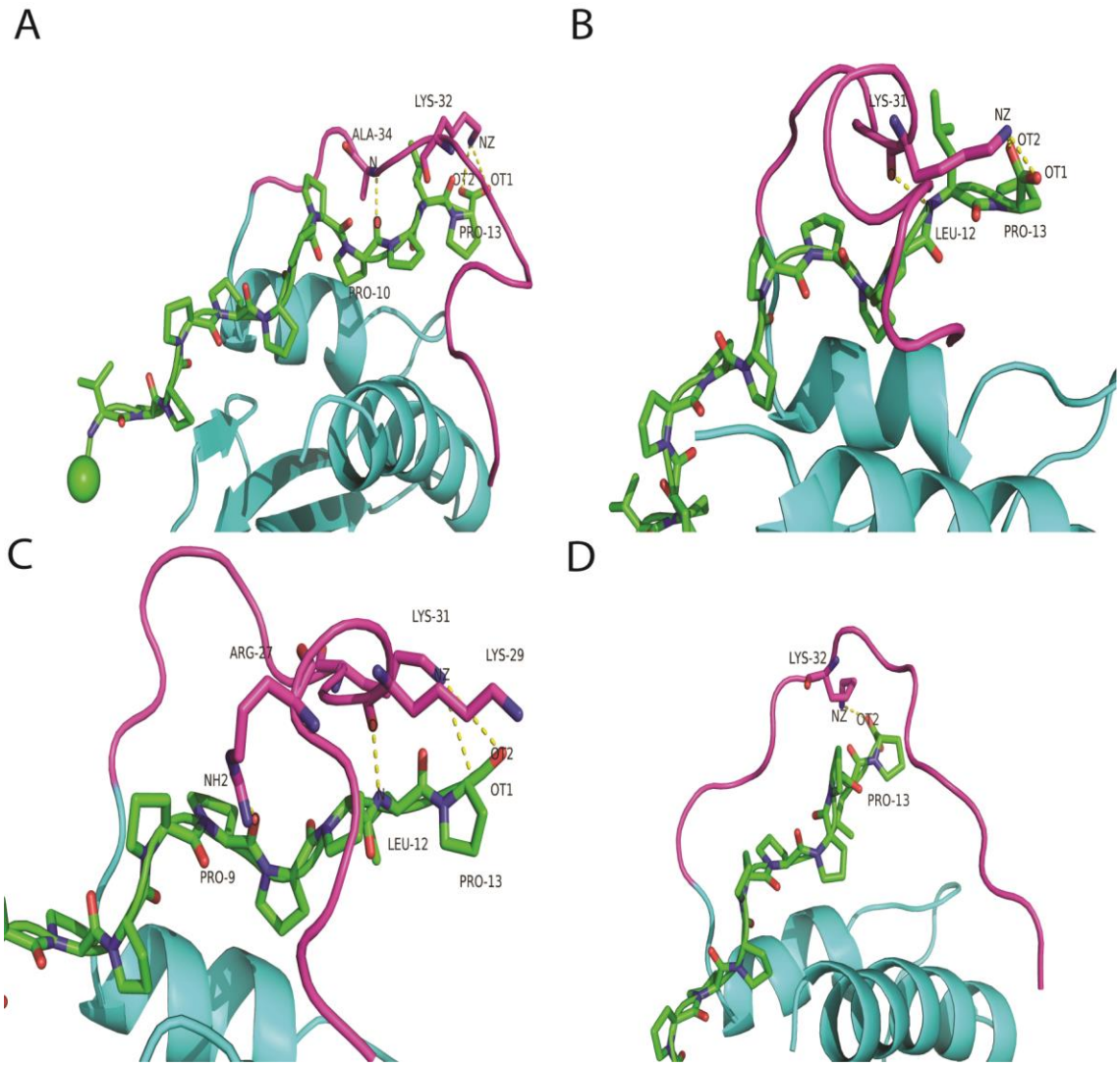

Figure 5.6. Binding mode analysis of PRF3Δ22 variants in complex with PolyP. A. AtPRF3Δ22SM-PolyP cluster1. B. AtPRF3Δ22SM3RtoA-PolyP cluster 1. C. AtPRF3Δ22-PolyP cluster1. D. AtPRF3Δ22-PolyP cluster2. The AtPRF3 Profilin core shown as cyan cartoon. AtPRF3 NTE shown as magenta cartoon. Residues of the AtPRF3 NTE identified to be involved in biomolecular interactions shown as magenta sticks. PolyP shown as green sticks.

Table 5.2 Calculated contact occupancies of identified biomolecular interactions in AtPRF3Δ22-NTE variants complexed with PolyP

Binding Mode	PRF3	PolyP	Type	Occupancy (%)
AtPRF3Δ22SM-PolyP-Cluster1	K32-NZ	P13-OT1	Electrostatic	54.6
	K32-NZ	P13-OT2	Electrostatic	54.7
	A34-N	P10-O	Hydrogen Bond	25.3
AtPRF3Δ22SM3RA-PolyP-Cluster1	K31-O	L12-N	Hydrogen Bond	26.6
	K31-NZ	P13-OT1	Electrostatic	55.8
	K31-NZ	P13-OT2	Electrostatic	55.2
AtPRF3Δ22-PolyP-Cluster1	R27-NH2	P9-O	Hydrogen Bond	55.8
	K29-O	L12-N	Hydrogen Bond	28.3
	K31-NZ	P13-OT1	Electrostatic	33.3
	K31-NZ	P13-OT2	Electrostatic	32.9
AtPRF3Δ22-PolyP-Cluster2	K32-NZ	P13-OT2	Electrostatic	50.4

Table 5.3. Binding Free energy Calculations using MM/PBSA for AtPRF3 variants in complex with PolyP. Units in KJ/mol

Cluster	Percentage Population (%)	Evdw	Eelec	Epolar	Eapolar	Ebinding
AtPRF3Δ22SM-PolyP-Cluster1	82.63	-301.306 ± 0.675	-553.927 ± 2.382	605.557 ± 2.947	-35.945 ± 0.079	-285.674 ± 1.804
AtPRF3Δ22SM3RA-PolyP-Cluster1	32.05	-276.586 ± 0.861	-575.155 ± 3.003	582.271 ± 3.605	-33.819 ± 0.095	-303.033 ± 1.551
AtPRF3Δ22-PolyP-Cluster1	76.64	-289.099 ± 0.591	-573.063 ± 2.513	606.003 ± 3.075	-34.373 ± 0.055	-290.624 ± 1.140
AtPRF3Δ22-PolyP-Cluster2	18.62	-284.062 ± 0.942	-490.705 ± 4.000	508.456 ± 4.860	-34.690 ± 0.115	-300.987 ± 2.043
AtPRF3Δ37-PolyP-Cluster1	100	-200.328 ± 0.346	-287.724 ± 1.276	259.269 ± 1.697	-22.582 ± 0.042	-251.381 ± 0.957

5.3.4 Alternative model of AtPRF3Δ22SM-PolyP complex with NTE modelled in partial helical conformation

Our model was of the NTE initiated from disordered conformation. The PSIPRED server predicted a short helical region of the NTE from residues S26 to K32 in AtPRF3 (Fig. 5.7). We hence modelled the NTE of the AtPRF3Δ22SM with an alternative conformation of a partial helix as detailed by the PSIPRED[172] server. We performed similar Classical Molecular Dynamics simulations, clustering and binding free energy calculations with a protocol similar as that performed for the other AtPRF3Δ22-NTE variants, to understand the effect of modelling the AtPRF3Δ22SM NTE in a partial helical conformation on the binding towards PolyP. We

identified three clusters of AtPRF3Δ22SM-PolyP conformations in which the NTE was modelled in partial helical conformations (Fig. 5.8). The calculated binding free energies AtPRF3Δ22SM-PolyP complexes with the NTE modelled in partial helical conformation were in a similar order of magnitude (Table 5.4) as that of AtPRF3Δ22SM-PolyP complex with the NTE in a disordered conformation (Table 5.3). Because of the limitations of classical molecular dynamics sampling, we were unable to observe the interconversion between helical and disordered states of the NTE. We have constructed an alternative model of AtPRF3Δ22SM-PolyP complex with the possibility of the NTE adopting partial helical conformation.

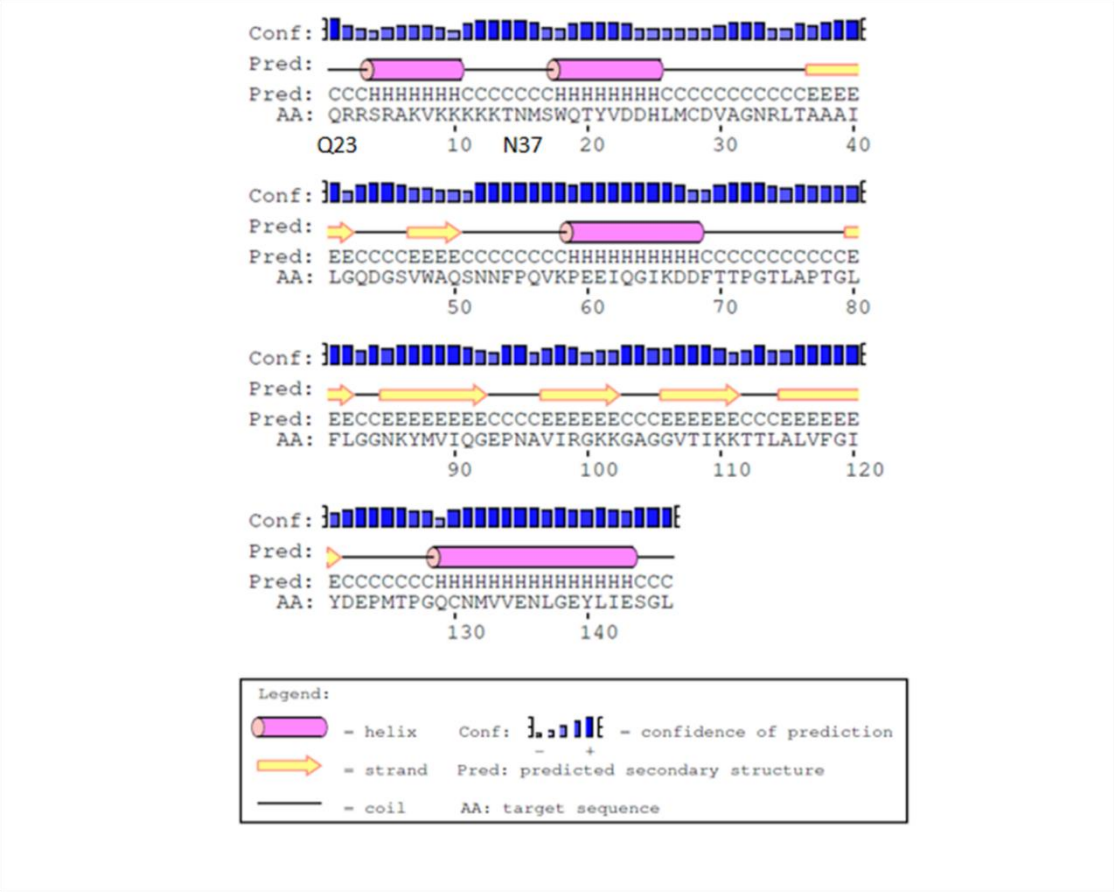

Figure 5.7 PSIPRED server prediction of AtPRF3 secondary structure

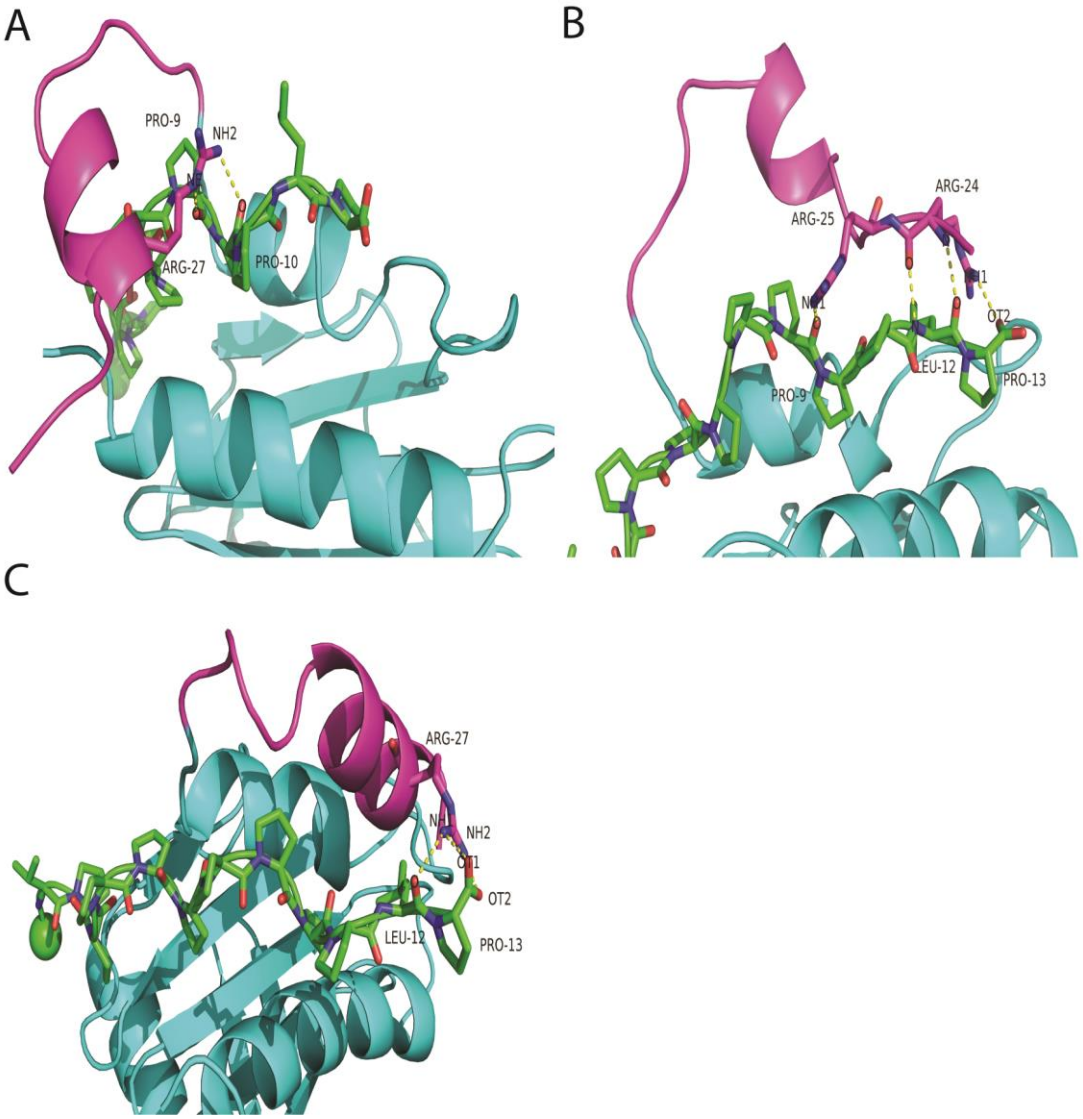

Figure 5.8. Binding pose analysis AtPRF3Δ22SM-PolyP complex with NTE modelled in a partial helical conformation. A. AtPRF3Δ22SM-PolyP cluster1. B. AtPRF3Δ22SM-PolyP cluster2. C. AtPRF3Δ22SM-PolyP cluster3.

Table 5.4 Binding Free energy Calculations using MM/PBSA for PolyP bound AtPRF3Δ22SM with NTE partial helical conformation. Units in KJ/mol

Cluster	Percentage Population (%)	Evdw	Eelec	Epolar	Eapolar	Ebinding
PRF3-Δ22-PolyP-Cluster1	40.9	-311.155 ± 0.823	-263.624 ± 2.892	346.367 ±2.788	-34.482 ±0.092	-262.868 ±1.521
PRF3-Δ22-PolyP-Cluster2	39.6	-325.313 ± 1.163	-549.880 ±2.987	593.162 ±3.478	-37.863 ±0.119	-320.011 ±1.450
PRF3-Δ22-PolyP-Cluster3	18.3	-250.101 ± 1.089	-562.917 ±5.350	515.499 ±5.517	-30.483 ±0.115	-327.959 ±2.693
PRF3-Δ37-PolyP-Cluster1	100	-200.328 ± 0.346	-287.724 ±1.276	259.269 ±1.697	-22.582 ±0.042	-251.381 ±0.957

5.3.5 Perspective of Simulation Findings

Intrinsically Disordered Regions play important roles by regulating different cellular activities. The conformational flexibility and diversity allows them to function in fine-tuning cellular process, such as direct protein-protein interactions[16]. Our study has demonstrated the conformational plasticity of the AtPRF3 and its role in enhancing binding towards PolyP. We showed that the NTE binds to PolyP in an adaptive manner that enables different binding modes to maintain the interaction between AtPRF3 and PolyP, as a result of its intrinsically disordered nature. Our simulation generated atomistic models of the AtPRF3-PolyP complexes provide an important tool to rationalize the experimental results of similar binding affinities of different AtPRF3 NTE variants towards PolyP. The binding of profilins to PolyP is an important aspect of formin mediated actin assembly. Such adaptive conformations of the AtPRF3 NTE towards PolyP is hypothesised to allow the maintenance of a continuous interaction between AtPRF3 and formin for the effective control of actin polymerization in plant innate immunity[169].

5.4 Conclusion and future directions

In this project, we used molecular dynamics simulations to study the conformational flexibility of the AtPRF3 NTE and its importance in enhancing binding towards PolyP. The NTE in crystal structure of AtPRF3Δ30SM was found to adopt a "closed" conformation, which occludes the PolyP binding pocket. Our H-REMD simulations and free energy surface as a function of the dihedral angle formed by C-alpha carbons of AtPRF3 residues 36-39 against the center of the mass distance between the NTE and the C-terminal Helix revealed the conformational flexibility of the NTE. The NTE is able to adopt an either "closed" or "open" conformations, where an "open" conformation reveals an accessible pocket for PolyP binding. Furthermore, we have used molecular dynamics simulations, geometric clustering, and binding free energy calculations to study the interactions between AtPRF3Δ22 variants and PolyP. We propose a model that AtPRF3 NTE interacts with PolyP via dynamic, flexible, and non-specific binding modes. This is supported by experimental biochemistry microscale thermophoresis assays, whereby marginal differences were observed between the binding affinities of AtPRF3Δ22 variants towards PolyP. This is also consistent with the nature of intrinsically disordered regions, which are characterized by conformational flexibility and diversity[14].

There are several future directions to this project. The existence of the NTE is not exclusive to AtPRF3. Sun et al showed that such an extension exists in different species, though there is little sequence conservation[166]. Such Intrinsically Disordered Region-mediated interactions for protein-protein interactions are worthy of future investigations from other protein systems. Also, a common characteristic of intrinsically disordered regions is the presence of Post Translational Modifications (PTMs)[173], which may potentially alter the charges of residues[17]. The effect of PTMs on the conformational flexibility of the NTE can be investigated in a future work.

6. Conclusions

In this thesis, we applied Molecular Dynamics simulations to understand the themes of Biomolecular folding and recognition. These are two processes which proteins undergo in order to perform their biological functions. The various simulation findings and biological implications for each project have been summarized and concluded at the end of each chapter.

MD simulations have been shown to be an important tool in the study of biological phenomenon. Structural biology experiments such as X-Ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy are filled with their inherent challenges and limitations, especially for conformationally flexible systems, such as Intrinsically Disordered Proteins and Regions (IDPs/IDRs). The atomistic resolution of MD simulations allows the construction of binding mode models, which can provide insight into the biomolecular interactions between biomolecules. This is evident for Project 2, in which we identified binding modes of the Aβ40 peptide towards L-PGDS. The ability of MD simulations to generate conformational ensembles at specific conditions (eg. temperature and pressure) provide the ability to explain and rationalize observed biological phenomenon. Many biological phenomena cannot be explained by a single static conformation generated by experimental structural biology methods. This is evident in Project 1, where conformational differences between the HAMLET Alpha1 peptide in its oleate bound form and its apo form are rationalized by MD simulations. This is also observed for Project 3, where the AtPRF3 was observed to adopt “open” and “closed” conformations in it’s Apo form, and adopt multiple adaptive non-specific binding modes toward PolyP. Throughout this thesis, we have shown how our simulation findings are complementary to experimental results to provide insight into the understanding of biological phenomena. Where possible, we have shown agreement between our simulation data and experimental measurements.

Biomolecular interactions are important for biomolecular folding and recognition. The array of biomolecular interactions such as electrostatic and aromatic interactions confer specificity for biomolecular recognition. This can be observed in our binding studies of Aβ40 towards the L-PGDS peptide. In addition, biomolecular recognition may be non-specific in nature. This can be observed in conformationally flexible biomolecules such as IDPs/IDRs. IDPs and IDRs have been evolutionarily optimized to be flexible in their conformations, to allow it to bind to multiple binding partners in a non-specific manner. In project 3, we show that the Arabidopsis thaliana Profilin N-Terminal extension (AtPRF3 NTE) has the ability to adopt multiple, non-specific binding modes in its interaction with poly-proline. Also, in Project 1, the interactions between oleate molecules and the Alpha1 peptide are shown to be non-specific, given the non-specific nature of hydrophobic interactions.

The conformational sampling of biomolecules, particularly those of IDPs/IDRs remains a computational challenge because of their conformationally flexible nature. With the advances in computational hardware, algorithmic and force field developments, it is expected that we would be able to increase sampling over longer timescales and larger system sizes. This would in turn contribute to our understanding of protein dynamics and protein biology.

List of manuscripts in submission/preparation

Ho, J.C.S., Nadeem, A., Kandiyal, P.S., *Ng, J.T.Y.*, Puthia,M.,Tran, T.H., Filenko,N., Hastings,A., Paul, S., Butler, D., Pimpalwar, N.,Tran, T.H., Fortunati, D.L. ,Hansen, J.S., Mu, Y.G., Mok, K.H., and Svanborg, C. A novel, peptide-based approach for targeting and killing tumor cells with greater precision. (Submitted to Cancer Research).

References

1. Forbes, R.M., A.R. Cooper, and H.H. Mitchell, *The composition of the adult human body as determined by chemical analysis.* J. Biol. Chem. , 1953. **203**: p. 359-366.
2. Alberts, B., et al., *Molecular Biology of the Cell. 5th edition*. 2008, New York: Garland Science.
3. Soto, C., *Protein misfolding and neurodegeneration* JAMA Neurology, 2008. **65**(2): p. 184-189.
4. Fernández-Medarde, A. and E. Santos, *Ras in cancer and developmental diseases.* Genes Cancer, 2011. **2**(3): p. 344–358.
5. Lodish, H., et al., *Molecular Cell Biology Fifth Edition*. 2003: W. H. Freeman.
6. Eisenberg, D., *The discovery of the α-helix and β-sheet, the principal structural features of proteins.* PNAS, 2003. **100**(20): p. 11207-11210.
7. Perutz, M.F., et al., *Structure of haemoglobin. A three-dimensional Fourier synthesis at 5.5Å resolution, obtained by X-ray analysis.* Nature, 1960. **185**: p. 416–422.
8. Haber, E. and C.B. Anfinsen, *Side-chain interactions governing the pairing of half-cystine residues in ribonuclease.* J Biol Chem, 1962. **237**: p. 1839-44.
9. Anfinsen, C.B., *Principles that govern the folding of protein chains.* Science, 1973. **181**(4096): p. 223-30.
10. Levinthal, C., *How to Fold Graciously.* Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois, 1969: p. 22-24.
11. Dill, K. and H.S. Chan, *From Levinthal to pathways to funnels.* Nat. Struct. Biol, 1997. **4**(1): p. 10-19.
12. Schug, A. and J.N. Onuchic, *From protein folding to protein function and biomolecular binding by energy landscape theory. .* Current Opinion in Pharmacology, 2010. **10**: p. 709-714.
13. Iadanza, M.G., et al., *A new era for understanding amyloid structures and disease.* Nature reviews. Molecular Cell Biology. , 2018. **19**.
14. Uversky, V.N., *The Mysterious Unfoldome: Structureless, Underappreciated, Yet Vital Part of Any Given Proteome. .* Journal of Biomedicine and Biotechnology 2009. **2010**(568068).
15. Oldfield, C.J., et al., *Comparing and combining predictors of mostly disordered proteins.* Biochemistry, 2005. **44**: p. 1989-2000.
16. Uversky, V.N., *Functional roles of transiently and intrinsically disordered regions within proteins. .* FEBS Journal 2015. **282**: p. 1182–1189
17. Uversky, V.N., *Dancing Protein Clouds: The Strange Biology and Chaotic Physics of Intrinsically Disordered Proteins. .* THE JOURNAL OF BIOLOGICAL CHEMISTRY 2016. **291**(13): p. 6681-6688.
18. Papoian, G.A., *Proteins with Weakly Funneled Energy Landscapes Challenge the Classical Structure-Function Paradigm.* Proc. Natl. Acad. Sci. , 2008. **105**: p. 14237–14238.
19. Ho, J.C.S., A. Nadeem, and C. Svanborg, *HAMLET- A protein-lipid complex with broad tumoricidal activity.* Biochemical and Biophysical Research Communications, 2017. **482**(2017): p. 454-458.
20. Permyakov, E.A. and L.J. Berliner, *alpha-Lactalbumin: structure and function.* FEBS Lett, 2000. **473**(3): p. 269-74.

21. Pettersson-Kastberg, J., et al., *Can misfolded proteins be beneficial? The HAMLET case.* Ann Med, 2009. **41**(3): p. 162-76.
22. Fischer, E., *Einfluss der Configuration auf die Wirkung der Enzyme.* Ber. Dtsch. Chem. Ges, 1894. **27**: p. 2985–2993.
23. Koshland, D., *Application of a theory of enzyme specificity to protein synthesis.* Proc. Natl. Acad. Sci. U.S.A, 1958. **44**: p. 98-104.
24. Ma, B., et al., *Folding funnels and binding mechanisms.* Protein Eng, 1999. **12**(9): p. 713-720.
25. Fenwick, R.B., S. Esteban-Martín, and X. Salvatella, *Understanding biomolecular motion, recognition, and allostery by use of conformational ensembles.* Eur Biophys J, 2011. **40**(12): p. 1339-1355.
26. Boehr, D.D., R. Nussinov, and P.E. Wright, *The role of dynamic conformational ensembles in biomolecular recognition.* Nat Chem Biol, 2009. **5**(11): p. 789–796.
27. Silva, D.A., et al., *A Role for Both Conformational Selection and Induced Fit in Ligand Binding by the LAO Protein.* PLoS Comput Biol 2011. **7(5): e1002054**.
28. Tanford, C., *The hydrophobic effect: formation of micelles and biological membranes*. 1973, New York: Wiley.
29. Southall, N.T., K.A. Dill, and A.D.J. Haymet, *A View of the Hydrophobic Effect.* J. Phys. Chem. B, 2002. **106**(3): p. 521-533.
30. Kahn, K. and K.W. Plaxco, *Principles of Biomolecular Recognition*, in *Recognition Receptors in Biosensors*, M. Zourob, Editor. 2010, Springer Science Business Media. p. 3-45.
31. Freire, E., *Do enthalpy and entropy distinguish first in class from best in class? .* Drug Discov Today, 2008. **13**((19-20)): p. 869–874.
32. Bissantz, C., B. Kuhn, and M. Stahl, *A Medicinal Chemist's Guide to Molecular Interactions.* J. Med. Chem, 2010. **53**: p. 5061-5084.
33. Hollingsworth, S.A. and R.O. Dror, *Molecular Dynamics Simulation for All.* Neuron, 2018. **99**: p. 1129-1143.
34. Hospital, A., et al., *Molecular dynamics simulations: advances and applications.* Adv Appl Bioinform Chem, 2015. **8**: p. 37-47.
35. Adcock, S.A. and J.A. McCammon, *Molecular Dynamics: Survey of Methods for Simulating the Activity of Protein.* Chem Rev, 2006. **106**(5): p. 1589–1615.
36. McCammon, J.A., B.R. Gelin, and M. Karplus, *Dynamics of folded proteins.* Nature, 1977. **267**: p. 585-590.
37. Shaw, D.E., et al., *Atomic-level characterization of the structural dynamics of proteins.* Science, 2010. **330**: p. 341-346.
38. Lindorff-Larsen, K., et al., *Picosecond to Millisecond Structural Dynamics in Human Ubiquitin* J. Phys. Chem. B 2016. **120**: p. 8313-8320.
39. Shaw, D.E., et al. *Millisecond-Scale Molecular Dynamics Simulations on Anton*. in *the Conference on High Performance Computing, Networking, Storage and Analysis (SC09)*. 2009. New York.
40. Li, B., et al., *Finding the needle in the haystack: towards solving the proteinfolding problem computationally.* Crit Rev Biochem Mol Biol, 2018. **53**(1): p. 1-28.
41. Plattner, N., et al., *Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling* NATURE CHEMISTRY 2017. **9**: p. 1005-1011.

42. Dror, R.O., et al., *Biomolecular Simulation: A Computational Microscope for Molecular Biology.* Ann Rev Biophys, 2012. **41**: p. 429-452.
43. Davis, A.M., S.J. Teague, and G.J. Kleywegt, *Application and Limitations of X-ray Crystallographic Data in Structure-Based Ligand and Drug Design.* Angew. Chem. Int. Ed., 2003. **42**: p. 2718–2736
44. Diasa, D.M. and A. Ciulli, *NMR approaches in structure-based lead discovery: Recent developments and new frontiers for targeting multi-protein complexes.* Prog Biophys Mol Biol, 2014. **116**(2-3): p. 101-112.
45. Goh, C.S., D. Milburn, and M. Gerstein, *Conformational changes associated with protein-protein interactions.* Curr Opin Struct Biol, 2004. **14**(1): p. 104-109.
46. Cavasotto, C.N., N.S. Adler, and M.G. Aucar, *Quantum Chemical Approaches in Structure-Based Virtual Screening and Lead Optimization.* Front Chem, 2018. **6**(188).
47. Cramer, C.J., *Essentials of computational chemistry: Theories and models*. 2006: J. Wiley & Sons.
48. *CPMD*. 1990-2019, IBM Corp.
49. Hutter, J., et al., *cp2k: atomistic simulations of condensed matter systems.* WIREs Comput Mol Sci, 2014. **4**: p. 15-25.
50. González, M.A., *Force fields and molecular dynamics simulations.* Collection SFN, 2011. **12**: p. 169-200.
51. MacKerell, A.D., et al., *All-atom empirical potential for molecular modeling and dynamics studies of proteins.* J Phys Chem B, 1998. **102**(18): p. 3586-616.
52. Cornell, W.D., et al., *A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules.* Journal of the American Chemical Society, 1995. **117**(19): p. 5179-5197.
53. Oostenbrink, C., et al., *A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6.* J Comput Chem., 2004. **25**(13): p. 1656-1676.
54. Jorgensen, W.L., D.S. Maxwell, and J. Tirado-Rives, *Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids.* Journal of the American Chemical Society, 1996. **118**(45): p. 11225-11236.
55. Durrant, J.D. and J.A. McCammon, *Molecular dynamics simulations and drug discovery.* BMC Biol, 2011. **9**: p. 71.
56. Huang, J., et al., *CHARMM36m: an improved force field for folded and intrinsically disordered proteins.* 2017. **14**(1): p. 71-73.
57. Allen, M.P. and D.J. Tildesley, *Computer simulation of liquids.* Clarendon Press, Oxford, 1987.
58. Van Der Spoel, D., et al., *GROMACS: fast, flexible, and free.* J Comput Chem, 2005. **26**(16): p. 1701-18.
59. Spiwok, V., Z. Sucur, and P. Hosek, *Enhanced sampling techniques in biomolecular simulations.* Biotechnology Advances, 2015. **33**(6): p. 1130-1140.
60. Bernardi, R.C., M.C.R. Melo, and K. Schulten, *Enhanced Sampling Techniques in Molecular Dynamics Simulations of Biological Systems.* Biochim Biophys Acta, 2015. **1850**(5): p. 872–877.
61. Ostermeir, K. and M. Zacharias, *Advanced replica-exchange sampling to study the flexibility and plasticity of peptides and proteins.* Biochim Biophys Acta, 2013. **1834**(5): p. 847-53.

62. Kannan, S. and M. Zacharias, *Enhanced sampling of peptide and protein conformations using replica exchange simulations with a peptide backbone biasing-potential.* Proteins, 2007. **66**(3): p. 697-706.
63. Hritz, J. and C. Oostenbrink, *Hamiltonian replica exchange molecular dynamics using soft-core interactions.* J Chem Phys, 2008. **128**(14): p. 144121.
64. Wang, L., R.A. Friesner, and B.J. Berne, *Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2).* J Phys Chem B, 2011. **115**(30): p. 9431-8.
65. Liu, P., et al., *Replica exchange with solute tempering: A method for sampling biological systems in explicit water.* Proc Natl Acad Sci U S A, 2005. **102**(39): p. 13749–13754.
66. Bussi, G., *Hamiltonian replica exchange in GROMACS: a flexible implementation.* Molecular Physics, 2014. **112**(3-4): p. 379-384.
67. Terakawa, T., T. Kameda, and S. Takada, *On easy implementation of a variant of the replica exchange with solute tempering in GROMACS.* J Comput Chem, 2011. **32**(7): p. 1228-1234.
68. Laio, A. and F.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, 2008. **71**(12): p. 126601.
69. Pfaendtner, J., *Metadynamics to Enhance Sampling in Biomolecular Simulations.* Methods Mol Biol, 2019. **2022**: p. 179-200.
70. Genheden, S. and U. Ryde, *The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities.* Expert Opin Drug Discov, 2015. **10**(5): p. 449-61.
71. Gao, C., M.S. Park, and H.A. Stern, *Accounting for ligand conformational restrictions in calculations of protein--ligand binding affinities.* Biophys J, 2010. **98**: p. 901-910.
72. Hou, T., et al., *Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations.* J. Chem. Inf. Model. , 2011. **51**: p. 69-82.
73. Homeyer, N. and H. Gohlke, *Free energy calculations by the molecular mechanics Poisson--Boltzmann surface area method.* Mol Inf 2012. **31**: p. 114-122.
74. Arodola, O.A. and M.E.S. Soliman, *Quantum mechanics implementation in drug-design workflows: does it really help? .* Drug Des Devel Ther, 2017. **11**: p. 2551–2564.
75. Nolan, T., N. Singh, and C.R. McCurdy, *Ligand macromolecule interactions: theoretical principles of molecular recognition.* Methods Mol Biol, 2009. **572**: p. 13-29.
76. Piovesan, D., G. Minervini, and S.C. Tosatto, *The RING 2.0 web server for high quality residue interaction networks.* Nucleic Acids Res, 2016. **44**: p. W367–W374.
77. Nicholls, A., et al., *Predicting small-molecule solvation free energies: an informal blind test for computational chemistry.* J Med Chem, 2008. **51**(4): p. 769-779.
78. Dong, F., B. Olsen, and N.A. Baker, *Computational methods for biomolecular electrostatics.* Methods Cell Biol, 2008. **84**: p. 843-870.
79. Israelachvili, J.N., *Intermolecular and Surface Forces*. 2011, Burlington, MA: Academic Press.
80. Krause, H., B. Ernstberger, and H.J. Neusser, *Binding energies of small benzene clusters.* Chem. Phys. Lett, 1991. **184**: p. 411-417.
81. Dougherty, D.A., *Cation/pi interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp.* Science, 1996. **271**: p. 163-168.

82. Dougherty, D.A., *The Cation-π Interaction.* Acc Chem Res, 2013. **46**(4): p. 885-893.
83. Sussman, J.L., et al., *Atomic structure of acetylcholinesterase from Torpedo californica: a prototypic acetylcholine-binding protein.* Science, 1991. **253**: p. 872-879.
84. Tsuzuki, S., et al., *The Magnitude of the CH/π Interaction between Benzene and Some Model Hydrocarbons.* J. Am. Chem. Soc, 2000. **122**(15): p. 3746-3753.
85. Hakansson, A., et al., *Apoptosis induced by a human milk protein.* Proc Natl Acad Sci U S A 1995. **92**: p. 8064-8068.
86. Acharya, K.R., et al., *Crystal structure of human alpha-lactalbumin at 1.7 A resolution.* J. Mol. Biol., 1991. **221**.
87. Svensson, M., et al., *Molecular characterization of α-lactalbumin folding variants that induce apoptosis in tumor cells.* J. Biol. Chem., 1999. **274**(10): p. 6388-6396.
88. Barbara, C. and M.D. Perez, *Interaction of α-lactalbumin with lipids and possible implications for its emulsifying properties: a review.* Int. Dairy J. , 2011. **21**: p. 727-741.
89. Mossberg, A.K., et al., *Structure and function of human alpha-lactalbumin made lethal to tumor cells (HAMLET)-type complexes.* Febs j, 2010. **277**(22): p. 4614-25.
90. Svensson, M., et al., *Alpha-lactalbumin unfolding is not sufficient to cause apoptosis, but is required for the conversion to HAMLET (human alpha-lactalbumin made lethal to tumor cells).* Protein Sci, 2003. **12**(12): p. 2794-804.
91. Pettersson-Kastberg, J., et al., *alpha-Lactalbumin, engineered to be nonnative and inactive, kills tumor cells when in complex with oleic acid: a new biological function resulting from partial unfolding.* J Mol Biol, 2009. **394**(5): p. 994-1010.
92. Svensson, M., et al., *Conversion of alpha-lactalbumin to a protein inducing apoptosis.* Proc Natl Acad Sci U S A, 2000. **97**(8): p. 4221-6.
93. Nadeem, A., et al., *Protein receptor-independent plasma membrane remodeling by HAMLET: a tumoricidal protein-lipid complex.* Sci Rep, 2015. **5**(16432).
94. Storm, P., et al., *A unifying mechanism for cancer cell death through ion channel activation by HAMLET.* PLoS One, 2013. **8**(3): p. e58578.
95. Hallgren, O., et al., *HAMLET triggers apoptosis but tumor cell death is independent of caspases, Bcl-2 and p53.* Apoptosis, 2006. **11**(2): p. 221-33.
96. Ho, J.C., et al., *Targeting of nucleotide-binding proteins by HAMLET- a conserved tumor cell death mechanism.* Oncogene, 2016. **35**(7): p. 897-907.
97. Gustafsson, L., et al., *Changes in proteasome structure and function caused by HAMLET in tumor cells.* PLoS One, 2009. **4**(4): p. e5229.
98. Kohler, C., et al., *A folding variant of human alpha-lactalbumin induces mitochondrial permeability transition in isolated mitochondria.* Eur J Biochem, 2001. **268**(1): p. 186-91.
99. Svanborg, C., et al., *HAMLET kills tumor cells by an apoptosis-like mechanism—cellular, molecular, and therapeutic aspects.* Adv. Cancer Res, 2003. **88**: p. 1-29.
100. Fischer, W., et al., *Human α-lactalbumin made lethal to tumor cells (HAMLET) kills human glioblastoma cells in brain xenografts by an apoptosis-like mechanism and prolongs survival.* Cancer Res, 2004. **64**: p. 2105–2112.
101. Gustafsson, L., et al., *Treatment of skin papillomas with topical α-lactalbumin-oleic acid.* N. Engl. J. Med. , 2004. **350**: p. 2663–2672.

102. Mossberg, A.K., et al., *Bladder cancers respond to intravesical instillation of HAMLET (human α-lactalbumin made lethal to tumor cells). .* Int. J. Cancer 2007. **121**: p. 1352–1359.
103. Ho, C.S., et al., *Low resolution solution structure of HAMLET and the importance of its alpha-domains in tumoricidal activity.* PLoS One, 2012. **7**(12): p. e53051.
104. Harata, K., Y. Abe, and M. Muraki, *Crystallographic evaluation of internal motion of Human alpha-lactalbumin refined by full-matrix least-squares method.* J.Mol.Biol, 1999. **287**: p. 347-358.
105. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water.* The Journal of chemical physics, 1983. **79**(2): p. 926-935.
106. Lindorff-Larsen, K., et al., *Improved side-chain torsion potentials for the Amber ff99SB protein force field.* Proteins, 2010. **78**(8): p. 1950-8.
107. *Accelrys Discovery Studio,*. Accelrys Software Inc: San Diego, CA, USA.
108. Wang, J., et al., *Development and testing of a general amber force field.* J Comput Chem, 2004. **25**(9): p. 1157-74.
109. Frisch, M.J., et al., *Gaussian 09*. 2009, Gaussian, Inc: Wallingford, CT, USA.
110. Bayly, C.I., et al., *A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model.* The Journal of Physical Chemistry, 1993. **97**(40): p. 10269-10280.
111. Hess, B., *P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation.* J Chem Theory Comput, 2008. **4**(1): p. 116-22.
112. Darden, T., D. York, and L. Pedersen, *Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems.* The Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.
113. Bussi, G., D. Donadio, and M. Parrinello, *Canonical sampling through velocity rescaling.* J Chem Phys, 2007. **126**(1): p. 014101.
114. Bonomi, M., et al., *PLUMED: A portable plugin for free-energy calculations with molecular dynamics.* Computer Physics Communications, 2009. **180**(10): p. 1961-1972.
115. Mu, Y., P.H. Nguyen, and G. Stock, *Energy landscape of a small peptide revealed by dihedral angle principal component analysis.* Proteins, 2005. **58**(1): p. 45-52.
116. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-637.
117. Levitt, M., *Conformational preferences of amino acids in globular proteins.* Biochemistry, 1978. **17**(4277-4285).
118. Imai, K. and S. Mitaku, *Mechanisms of secondary structure breakers in soluble proteins.* Biophysics (Nagoya-shi), 2005. **1**: p. 55-65.
119. Wider, G., *Structure determination of biological macromolecules in solution using nuclear magnetic resonance spectroscopy.* BioTechniques 2000. **29**: p. 1278-1294.
120. Pande, V.S., *Understanding protein folding using Markov state models.* Adv Exp Med Biol, 2014. **797**: p. 101-106.
121. Husic, B.E. and V.S. Pande, *Markov State Models: From an Art to a Science.* J Am Chem Soc, 2018. **140**(7): p. 2386-2396.
122. Shukla, D., et al., *Markov state models provide insights into dynamic modulation of protein function.* Acc Chem Res, 2015. **48**(2): p. 414-422.
123. Pande, V.S. *Folding@home*. 2010; Available from: https://foldingathome.org/.

124. Prince, M., et al., *The global prevalence of dementia: A systematic review and metaanalysis.* Alzheimer's Dementia, 2013. **9**: p. 63-75.
125. Association., A.s., *2016 Alzheimer's disease facts and figures.* Alzheimer's & Dementia, 2016. **12**(4): p. 459-509.
126. Duthey, B., *WHO report Alzheimer Disease and other Dementias. Background paper 6.11*. 2013, World Health Organization. p. 1-74.
127. Lee, V.M., M. Goedert, and J.Q. Trojanowski, *Neurodegenerative tauopathies.* Annu Rev Neurosci, 2001. **24**: p. 1121–1159.
128. Selkoe, D.J., *Alzheimer's disease: genes, proteins, and therapy.* Physiol Rev, 2001. **81**(2): p. 741-766.
129. Murphy, M.P. and H. Levine, *Alzheimer's Disease and the β-Amyloid Peptide.* J Alzheimers Dis. , 2010. **19**(1): p. 311.
130. Selkoe, D.J. and J. Hardy, *The amyloid hypothesis of Alzheimer's disease at 25 years.* EMBO Mol Med 2016. **8**: p. 595-608.
131. Karran, E. and B. De Strooper, *The amyloid cascade hypothesis: are we poised for success or failure? .* J. Neurochem, 2016. **139**(2): p. 237-252.
132. Chaari, A., *Molecular chaperones biochemistry and role in neurodegenerative diseases.* Int J Biol Macromol, 2019. **131**: p. 396-411.
133. Gestwicki, J.E. and D. Garza, *Protein quality control in neurodegenerative disease.* Prog Mol Biol Transl Sci, 2012. **107**: p. 327-353.
134. Wyatt, A.R., et al., *Roles of Extracellular Chaperones in Amyloidosis.* J. Mol. Biol. , 2012. **421**: p. 499-516.
135. Sharma, B. and K. Pervushin, *Neuroprotective Function of Non-Proteolytic Amyloid-β Chaperones in Alzheimer's Disease*. Amyloid Diseases, ed. D. Kurouski. 2019: IntechOpen.
136. Kanekiyo, T., et al., *Lipocalin-type prostaglandin D synthase/beta trace is a major amyloid beta-chaperone in human cerebrospinal fluid.* Proc Natl Acad Sci U S A, 2007. **104**(15): p. 6412-6417.
137. Kume , S., et al., *Systematic interaction analysis of human lipocalin-type prostaglandin D synthase with small lipophilic ligands.* Biochem J., 2012. **446**(2): p. 279-289.
138. Mase, M., et al., *Lipocalin-type prostaglandin D synthase (beta-trace) in cerebrospinal fluid: A useful marker for the diagnosis of normal pressure hydrocephalus.* Neuroscience Research, 2003. **47**(4): p. 455-459.
139. Puchades, M., et al., *Proteomic studies of potential cerebrospinal fluid protein markers for Alzheimer's disease.* Brain Res Mol Brain Res, 2003. **118**(1-2): p. 140-146.
140. Kumasaka, T., et al., *Structural basis of the catalytic mechanism operating in open-closed conformers of lipocalin type prostaglandin D synthase.* J. Biol. Chem. , 2009. **284**(33): p. 22344-22352.
141. Zhou, Y., et al., *Structure-function analysis of human l-prostaglandin D synthase bound with fatty acid molecules.* FASEB J., 2010. **24**(12): p. 4668-4677.
142. Lim, S.M., et al., *Structural and dynamic insights into substrate binding and catalysis of human lipocalin prostaglandin D synthase.* J Lipid Res, 2013. **54**(6): p. 1630-43.
143. Rauth, S., et al., *High-affinity Anticalins with aggregation-blocking activity directed against the Alzheimer beta-amyloid peptide.* Biochem J, 2016. **473**(11): p. 1563-78.
144. Liu, Y., et al., *Binding Modes of Teixobactin to Lipid II: Molecular Dynamics Study.* Sci Rep, 2017. **7**(1): p. 17197.

145. Pagadala, N.S., K. Syed, and J. Tuszynski, *Software for molecular docking: a review.* Biophys Rev, 2017. **9**(2): p. 91-102.
146. Kumari, R. and R. Kumar, *g_mmpbsa-a GROMACS tool for high-throughput MM-PBSA calculations.* J Chem Inf Model, 2014. **54**(7): p. 1951-62.
147. Kannaian, B., et al., *Abundant neuroprotective chaperone Lipocalin-type prostaglandin D synthase (L-PGDS) disassembles the Amyloid-β fibrils.* Sci Rep, 2019. **9**(1): p. 12579.
148. Williamson, M.P., *Using chemical shift perturbation to characterise ligand binding.* Progress in Nuclear Magnetic Resonance Spectroscopy, 2013. **73**: p. 1-16.
149. Williamson, M.P., *Using chemical shift perturbation to characterise ligand binding.* Prog Nucl Magn Reson Spectrosc, 2013. **73**: p. 1-16.
150. Powers, K.T., M.S. Gildenberg, and M.T. Washington, *Modeling Conformationally Flexible Proteins With X-ray Scattering and Molecular Simulations.* Comput Struct Biotechnol J, 2019. **17**: p. 570-578.
151. Gao, L.B., et al., *Alzheimer's Disease therapeutics: current and future therapies.* Minerva Med, 2016. **107**(2): p. 108-113.
152. Tiwari, S., et al., *Alzheimer's disease: pathogenesis, diagnostics, and therapeutics.* Int J Nanomedicine, 2019. **14**: p. 5541-5554.
153. Cui, Q. and M. Karplus, *Allostery and cooperativity revisited.* Protein Sci, 2008. **17**(8): p. 1295-1307.
154. Tsai, C. and R. Nussinov, *A Unified View of "How Allostery Works".* PLoS Comput Biol, 2014. **10**(2): p. e1003394.
155. Vilches Barro, A., et al., *Cytoskeleton Dynamics Are Necessary for Early Events of Lateral Root Initiation in Arabidopsis.* Curr Biol., 2019. **29**(15): p. 2443-2454.
156. Li, J. and C.J. Staiger, *Understanding Cytoskeletal Dynamics During the Plant Immune Response.* Annu Rev Phytopathol, 2018. **56**: p. 513-533.
157. Rottner, K., et al., *Actin assembly mechanisms at a glance.* Journal of Cell Science, 2017. **2017**(130): p. 3427-3435.
158. Carlsson, L., et al., *Actin polymerizability is influenced by profilin, a low-molecular weight protein in non-muscle cells.* J Mol Biol, 1977. **115**(3): p. 465–483.
159. Kang, F., D.L. Purich, and F.S. Southwick, *Profilin Promotes Barbed-end Actin Filament Assembly without Lowering the Critical Concentration.* The Journal of Biological Chemistry, 1999. **274**: p. 36963-36972.
160. Pring, M., A. Weber, and M.R. Bubb, *Profilin-actin complexes directly elongate actin filaments at the barbed end.* Biochemistry, 1992. **31**(6): p. 1827-1836.
161. Petrella, E.C., et al., *Structural requirements and thermodynamics of the interaction of proline peptides with profilin.* Biochemistry, 1996. **35**(51): p. 16535–16543.
162. Paul, A. and T. Pollard, *The Role of the FH1 Domain and Profilin in Formin-Mediated Actin-Filament Elongation and Nucleation.* Curr Biol, 2008. **18**(1): p. 9-19.
163. Higgs, H.N., *Formin proteins: a domain-based approach.* Trends Biochem Sci, 2005. **30**(6): p. 342–353.
164. Courtemanche, N., *Mechanisms of formin-mediated actin assembly and dynamics.* Biophys Rev, 2018. **10**(6): p. 1553–1569.
165. Gunning, P.W., et al., *The evolution of compositionally and functionally distinct actin filaments.* Journal of Cell Science, 2015. **128**: p. 2009-2019.
166. Sun, H., et al., *Profilin Negatively Regulates Formin-Mediated Actin Assembly to Modulate PAMP-Triggered Plant Immunity.* Curr Biol. , 2018. **28**(12): p. 1882-1895.

167. Mahoney, N.M., P.A. Janmey, and S.C. Almo, *Structure of the profilin-poly-L-proline complex involved in morphogenesis and cytoskeletal regulation.* Nat Struct Biol, 1997. **4**(11): p. 953-960.
168. Kursula, P., et al., *High-resolution Structural Analysis of Mammalian Profilin 2a Complex Formation with Two Physiological Ligands: The Formin Homology 1 Domain of mDia1 and the Proline-rich Domain of VASP* J. Mol. Biol., 2008. **375**: p. 270-290.
169. Qiao, Z., et al., *Structural and computational examination of the Arabidopsis profilin-Poly-P complex reveals mechanistic details in profilin-regulated actin assembly.* J Biol Chem, 2019. **294**(49): p. 18650-18661.
170. Mahoney, N.M., et al., *Profilin binds proline-rich ligands in two distinct amide backbone orientations.* Nature structural biology, 1999. **6**(7): p. 666-671.
171. Lopes, J.L., et al., *Distinct circular dichroism spectroscopic signatures of polyproline II and unordered secondary structures: applications in secondary structure analyses.* Protein Sci, 2014. **23**(12): p. 1765-1772.
172. Buchan, D.W.A., et al., *Scalable web services for the PSIPRED Protein Analysis Workbench.* Nucleic acids research, 2013. **41**: p. 349-357.
173. Bah, A. and J.D. Forman-Kay, *Modulation of Intrinsically Disordered Protein Function by Post-translational Modifications.* J Biol Chem, 2016. **291**(13): p. 6696–6705.