Molecular Modelling of Peptide Folding, Misfolding and Aggregation Phenomena

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Nevena Todorova

School of Applied Sciences
College of Science, Engineering and Health
RMIT University
August 2009
Declaration of Candidature

I certify that except where due acknowledgment has been made, the work is that of the candidate alone. This body of work has not been submitted previously, in whole or in part, to qualify for any other academic award. The content of this thesis is the results of work which has been carried out since the official commencement date of the approved research program. Any editorial work, paid or unpaid, carried out by a third part is acknowledged.

_________________________
Nevena Todorova
Acknowledgements

This project would not have existed without the contribution, support and guidance of many people who deserve to be acknowledged.

I wish to thank my supervisor, Prof. Irene Yarovskiy, for giving me the opportunity to work on this project. Her invaluable guidance and assistance throughout these years have made this research possible and also very enjoyable.

Дорогая Ирина,
Я никогда не забуду наши разговоры, смех, танцы и упражнения, это было замечательное время. Вы были не просто моим учителем, но и хорошим, близким другом. Спасибо за всё. С любовью, Невена.

I would also like to say a special thank you to my other supervisors, Dr. Sue Legge and Dr. Andrew Hung for their support, guidance and help during my research project. Your comments and suggestions that have made this thesis be a better piece of work.

Sincere thank you to Dr. Akin Budi (a.k.a Akinski) for accepting to read this thesis to the smallest detail as humanly possible. The hilarious comic printouts have made me look forward to your comments, thanks! :)

I would also like to acknowledge, Dr. Stefano Piana (Curtin University, Perth) and Fabrizio Marinelli (SISSA, Trieste) for their guidance and contribution to the work on the folding of insulin.

Geoff Howlett and group from the Bio21 Institute, Melbourne, thank you for the collaboration and experimental support you have provided to this project.

I would like to thank Dr. Herbert Treutlein (Cytopia Research Pty Ltd) for his contribution to the force field work.

Acknowledgments to the Australian Research Council and Cytopia Research Pty Ltd. for providing financial support on this project.

For providing all the computational facilities, I would like to acknowledge the Australian Partnership for Advanced Computing (APAC) and the Victorian Partnership for Advance Computing (VPAC).

To everyone in the Molecular Modelling and Simulations Group, thank you for making this project be a very enjoyable experience.
On a personal note, I would like to thank my dear friends for their support, understanding and encouragements over the past few years, with a special thank you to Zana for always being an e-mail or phone call away. Starbucks will never be the same!

I would like to say a special thank you to my partner Paul, for his love, support and understanding through these years, especially for his resilience over the last few months.

A very special thank you goes to my beautiful Sparky, you have brought so much joy and happiness to my life, and you are always there to cheer me up.

I would also like to thank my dearest grandma, who might be far away but never far from my heart. Your picture at my desk has given me strength to never give up.

And last but definitely not the least, I would like to thank and dedicate this thesis to my parents, for their love, understanding and never-ending support they always provide me. You are by backbone, my everything. I love and thank you greatly from the bottom of my heart. Бе сакам премногу, Невена.
Publication List

Peer-reviewed Publications


   “Computational investigation on the effects of mutation of the amyloidogenic apoC-II(60-70) monomers” (2009), in preparation.

   “Lipid concentration effects of the amyloidogenic apoC-II(60-70) peptide: Computational study” (2009), in preparation.
Conference Presentations

[1] N. Todorova
“Computational studies of peptide folding, misfolding and aggregation phenomena.”
The 12th Australian Molecular Modelling Conference, Gold Coast, Australia, 2009.

“Molecular modelling studies of peptide folding, misfolding and aggregation phenomena”
The 8th Melbourne Protein Group Symposium, Melbourne, Australia, 2009.

“Computational study on the effect of different lipid concentrations on the amyloidogenic peptide derived from apolipoprotein C-II.”

“Exploring the folding free energy landscape of insulin”
Biomolecular Interactions and Dynamics Symposium, Melbourne, 2009.
Award received for best Poster Presentation.

“Simulation study on the effect on stability of oxidation and mutation of amyloidogenic peptides derived from apolipoprotein C-II. Part 2: apoC-II (60-70)”

“BE-META: A new method for protein folding prediction, application to insulin”

“Exploring the folding free energy landscape of insulin”
Award received for best Poster Presentation.

“Computational methods for studying proteins”
Workshop: Interaction Energies and the Structure of Surfaces and Nano-Structures, RMIT, Melbourne, Australia, 2007

“Different force fields, one protein, do they give the same answer?”
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D-IR</td>
<td>Two-Dimensional Infrared Spectroscopy</td>
</tr>
<tr>
<td>AFM</td>
<td>Atomic Force Microscopy</td>
</tr>
<tr>
<td>APAC</td>
<td>Australian Partnership for Advanced Computing</td>
</tr>
<tr>
<td>ApoC-II</td>
<td>Apolipoprotein C-II</td>
</tr>
<tr>
<td>BE-META</td>
<td>Bias Exchange Metadynamics</td>
</tr>
<tr>
<td>CD</td>
<td>Circular Dichroism</td>
</tr>
<tr>
<td>D5PC</td>
<td>Dipentanoylphosphatidylcholine</td>
</tr>
<tr>
<td>DMD</td>
<td>Discrete Molecular Dynamics</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>DPI</td>
<td>Despentapeptide insulin</td>
</tr>
<tr>
<td>FRET</td>
<td>Fluorescence Resonance Energy Transfer</td>
</tr>
<tr>
<td>GROMACS</td>
<td>GROningen MAchine for Chemical Simulation</td>
</tr>
<tr>
<td>KMC</td>
<td>Kinetic Monte Carlo</td>
</tr>
<tr>
<td>MC</td>
<td>Monte Carlo</td>
</tr>
<tr>
<td>MCL</td>
<td>Markov Cluster analysis method</td>
</tr>
<tr>
<td>MD</td>
<td>Molecular dynamics</td>
</tr>
<tr>
<td>NAMD</td>
<td>Not Another Molecular Dynamics</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear Magnetic Resonance</td>
</tr>
<tr>
<td>NOE</td>
<td>Nuclear Overhauser Effect</td>
</tr>
<tr>
<td>NOESY</td>
<td>Nuclear Overhauser Effect Spectroscopy</td>
</tr>
<tr>
<td>NPT</td>
<td>Constants number of particles, pressure and temperature</td>
</tr>
</tbody>
</table>
NVT ............................................ Constant number of particles, volume and temperature
PBC .............................................................. Periodic boundary condition
PDB ............................................................. Protein Data Bank
RDF ...............................................................Radial Distribution Function
PEPCAT ............................................... Peptide Conformational Analysis Tool
REMD ..................................................... Replica Exchange Molecular Dynamics
RMSD ........................................................ Root Mean Square Displacement
RESP .............................................................. Restrained Electostatic Potential
SASA ............................................................. Solvent Accessible Surface Area
TFE ............................................................... Trifluoroethanol
ThT ................................................................. Theoflavin T
VMD ............................................................... Visual Molecular Dynamics
VPAC .............................................................. Victorian Partnership for Advanced Computing
WHAM ........................................................ Weighted Histogram Analysis Method
XRD ............................................................. X-ray Diffraction
# Table of Contents

Declaration of Candidature ....................................................................................................II
Acknowledgements ................................................................................................................ III
Publication List ........................................................................................................................ V
Abbreviations ........................................................................................................................ VII
Table of Contents ................................................................................................................... IX
List of Tables .......................................................................................................................... XIII
List of Figures .......................................................................................................................... XIV
Abstract .....................................................................................................................................1

1. Computational studies of protein folding and aggregation ..............................................3
   1.1 Overview ......................................................................................................................3
   1.2 Protein structure ..........................................................................................................4
   1.3 Protein folding .............................................................................................................11
      1.3.1 Studies of protein folding mechanisms ..............................................................11
      1.3.2 Protein folding kinetics ......................................................................................15
      1.3.3 Transition / intermediate states in protein folding ............................................16
      1.3.4 Unfolded or denatured states .............................................................................18
   1.4 Protein self-association and aggregation ....................................................................20
      1.4.1 Early history of amyloid diseases .....................................................................20
      1.4.2 Protein aggregation related diseases .................................................................21
      1.4.3 Structure of amyloid fibrils ...............................................................................23
      1.4.4 Mechanisms of fibril formation .........................................................................24
      1.4.5 Sequence and mutation effects on fibril formations ...........................................28
      1.4.6 Lipid effects on fibril formation .........................................................................29
      1.4.7 Oligomer stability ..............................................................................................30

2. Computational techniques for protein studies .................................................................32
   2.1 Overview ......................................................................................................................32
   2.2 Molecular modelling ..................................................................................................34
4.4.1 Secondary structure evolution .......................................................... 79
4.4.2 NOE energies and violations .......................................................... 84
4.4.3 Root mean square deviation (RMSD) ........................................... 91
4.4.4 Radius of gyration ................................................................. 93
4.4.5 Solvent accessible surface area (SASA) ....................................... 94
4.4.6 Cross-force field simulations ................................................... 95
4.5 Conclusions .............................................................................. 96

5. Exploring the folding free energy landscape of insulin ...................... 98
5.1 Overview ............................................................................ 98
5.2 Introduction .......................................................................... 99
5.3 Computational Details ........................................................... 100
5.3.1 Collective variables ............................................................ 102
5.3.2 Analytical method for biased statistics: Kinetic model construction ........................................................................ 103
5.4 Results and Discussions .......................................................... 104
5.4.1 Neutral replica analysis ........................................................ 105
5.4.2 Kinetic model construction .................................................. 108
5.4.3 Structural analysis of the folded state basin ............................ 114
5.4.4 Folding pathway of chain B of insulin ................................... 117
5.5 Conclusions ........................................................................ 119

6. Investigation on the effects of mutation of the amyloidogenic apoC-II(60-70) peptide ................................................................. 121
6.1 Overview ............................................................................ 121
6.2 Introduction .......................................................................... 122
6.3 Computational Details ........................................................... 125
6.4 Results and Discussions .......................................................... 127
6.4.1 Structural evolution analysis ................................................ 127
6.4.2 Cluster analysis ................................................................. 130
6.4.3 Root mean square deviation ................................................ 132
6.4.4 Radius of gyration ............................................................ 133
6.4.5 Side-chain interactions ........................................................ 135
6.4.6 Aromatic side-chain orientation ........................................... 137
6.4.7 Solvation properties ........................................................... 139
6.5 Conclusions ........................................................................ 142
7. Lipid concentration effects on the amyloidogenic apoC-II(60-70) peptide ..........144
   7.1 Overview ....................................................................................................................144
   7.2 Introduction ................................................................................................................145
   7.3 Computational Details ...............................................................................................147
   7.4 Results and Discussions ............................................................................................150
      7.4.1 Molecular dynamics: Lipid-peptide interactions .............................................150
      7.4.2 Molecular dynamics: Secondary structure evolution .......................................162
      7.4.3 Umbrella sampling: Free energy profiles .........................................................166
      7.4.4 Umbrella sampling: Lipid-peptide interactions ................................................167
   7.5 Conclusions ................................................................................................................170
8. Oligomers of apoC-II(60-70): stability & effect of mutation.................................173
   8.1 Overview .....................................................................................................................173
   8.2 Introduction ................................................................................................................174
   8.3 Computational Details ...............................................................................................175
   8.4 Results and Discussions ............................................................................................178
      8.4.1 Root mean square deviation and clustering analysis of wild-type apoC-II(60-
           70) oligomers ...........................................................................................................178
      8.4.2 Root mean square deviations of mutated apoC-II(60-70) oligomers ...............185
      8.4.3 Free energy profiles of wild-type and mutated apoC-II(60-70) dimers ............189
   8.5 Conclusions ................................................................................................................191
9. Future work ....................................................................................................................192
   9.1 Computational techniques and force field development .........................................192
   9.2 Theoretical investigation of fibril microcrystal structures .......................................193
   9.3 Lipids and apolipoprotein interactions in heart disease ..........................................193
   9.4 Protein self-assembly on surfaces, interfaces and nanoparticles .............................195
Appendix ............................................................................................................................197
References .........................................................................................................................208
List of Tables

Table 1.1 A list of diseases associated with the aggregation of proteins, adopted from the review article by Chiti and Dobson [101]. The name of the disease, the pathological protein, together with the size and native structure are specified.................................................................23

Table 3.1 Primary sequence of chain A and B of porcine insulin (PDB code 1ZNI). ..........61
Table 3.2 Specific sequence differences between human, porcine and bovine insulin.........62
Table 3.3 Primary sequence of apoC-II (PDB code 1SOH). The sequence coloured in grey represents the residues with undefined structure..........................69

Table 4.1 Average NOE interproton distance violations calculated for each force field simulation. The total average number of violations, as well as the severe violations larger than 1 Å from the upper bound value of the classified cross-peak intensities by Hawkins et al. [232] are indicated..................86

Table 4.2 Average radius of gyration (Rg) and its standard deviation, calculated for each simulation. ........................................................................................................93

Table 4.3 Average solvent accessible surface area (SASA) and standard deviations calculated for each simulation. ........................................................................94

Table 5.1 Thermodynamic properties of the basins found using the MCL algorithm with \( p = 1.08 \). All thermodynamic properties were calculated with respect to the native-like basins.................................................................111

Table 7.1 The nomenclature for the individual system, as well as their simulated times using: a) Molecular dynamics and b) Umbrella sampling method. The two digits in the label describe the ratio of lipid to peptide, and the letters \( c \) and \( h \) describe coil and helix starting structure, respectively........................................150

Table 8.1 System set-up listing the names given for each simulation, together with the simulation box dimensions, number of water molecules, counterions and total simulation time. .................................................................177
List of Figures

Figure 1.1 The general structure of an amino acid, with the amino group on the left, the carboxyl group on the right and the unique side-chain (R) bonded to the Ca atom. ..........................................................4

Figure 1.2 The condensation of two amino acids to form a peptide bond. .......................5

Figure 1.3 The amino acid sequence of a protein is called its primary structure. Different regions of the sequence form local regular secondary structures, such as α-helices or β-strands. The tertiary structure is formed by packing such structural elements into a globular unit or domain. The final conformation of a protein may contain several polypeptide chains arranged in a quaternary structure. ..................................................................7

Figure 1.4 Ramachandran diagram showing the three sterically allowed regions for common secondary structures of a protein. Lighter (yellow/green) colours represent the more favourable combinations of ψ-φ angle values. ......................8

Figure 1.5 Illustration of the hydrogen bonding patterns represented by dotted red lines in a) parallel; and b) anti-parallel β-sheets. .................................................................9

Figure 1.6 The β-hairpin structure; a) Inter-strand hydrogen bonding shown as black dotted lines (backbone shown only); b) Cartoon representation of β-hairpin. ......10

Figure 1.7 Cartoon representation of the different energy landscapes from a denatured conformation to the native conformation N, adopted from reference [16]. (a) A smooth energy landscape for fast folding protein; (b) A moat landscape, where folding must go through an obligatory intermediate; (c) A rugged energy landscape with kinetic traps; and (d) A golf course energy landscape in which folding is dominated by diffusional conformational search. ................13

Figure 1.8 Illustration of the free energy profile of a simple protein folding event, showing the unfolded (U), transition state (TS) and folded or native structure (N). ........................................................................................................17

Figure 1.9 Experimentally determined crystal structures of Aβ1-42 and Sup35 derived GNNQQNY peptides...........................................................................................................24

Figure 1.10 A schematic representation of some of the many conformational states adopted by polypeptide chains during folding, misfolding and aggregation.
The figure was reproduced from the detailed review presented by Chiti and Dobson [101].

Figure 2.1 Diagram of the accessible system size and time scale for different types of simulation studies.

Figure 2.2 A two-dimensional illustration of periodic boundary conditions.

Figure 2.3 Three-dimensional representation of a typical simulation box with solvent and protein model shown explicitly.

Figure 2.4 The real distribution of a system along a reaction coordinate “Distance” can be calculated using several umbrella sampling windows (red); after correcting for the bias potentials the free energy along the selected reaction coordinate is identified (purple).

Figure 2.5 In replica exchange molecular dynamics simulation several replica of the same system are simulated in parallel at different temperatures. Exchanges between neighbouring replicas are performed at selected times.

Figure 3.1 Representation from different angles of the porcine insulin monomer. Chain A is coloured blue, chain B is coloured red and the disulfide bonds between the Cys residues are coloured yellow.

Figure 3.2 An overlay of X-ray structures, showing experimentally observed states of insulin chain B.

Figure 3.3 The microcrystal structure of fibril forming peptide segment, VEALYL of insulin chain B (PDB code 2OMQ); a) A two-sheet motif of the smallest unit-cell of the VEALYL fibril structure, depicted in projection down the needle crystal axis; b) Cartoon representation of the crystalline seed; c) Longitudinal view of two-sheet fibril formation extending from the core seed structure represented in colour.

Figure 3.4 Ribbon representation of apoC-II(13-79) showing secondary structure and side-chains.

Figure 3.5 Amyloidogenic peptide ApoC-II(60-70).

Figure 4.1 Secondary structure evolution of insulin chain B. Simulations for each force field are labelled as described in the Computational Details section of this chapter. The secondary structure colour codes: magenta – α-helix, red – π-helix, cyan – turn, white – coil, yellow – extended conformation and green – bridge.
Figure 4.2 The reference structure of the two most populated clusters of each force field simulation, labelled as described in the Computational Details section of this chapter. The colouring of each protein is by its residue type, blue – basic, red – acidic, green – polar and silver – nonpolar. The hydrogen bonds of structure d2 are coloured purple...

Figure 4.3 NOE energy derived from the distance restraints of the sampled structures. The region below the line at $-41.5 \times 10^3$ a.u. represents the MD structures with NOE energy less than that of the starting structure. The plot shows a moving average of 250 ps...

Figure 4.4 Important protein conformations and their NOE distance violations. The satisfied distance restraints are shown in yellow and the violated distances in black. The colour scheme is the same as in Figure 4.1...

Figure 4.5 RMSD of the helical region of chain B, from residues Ser9 to Cys19, compared to the starting structure as a function of time for all systems. The plot shows a moving average of 250 ps...

Figure 4.6 Secondary structure evolution of insulin chain B; a) Starting structure taken from the last frame of simulation using GROMOSA6 force field was simulated using AMBER03; b) Starting structure taken from the last frame of simulation using CHARMM force field was simulated using OPLS...

Figure 5.1 The starting structure selected for this study on the folding mechanisms of insulin chain B...

Figure 5.2 The free energy of the 20 most populated clusters. The native-like cluster is shown as an inset of the figure. The structures showing the most typical features are represented on the right of the graph. The clusters are numbered in the order of free energy with ascending cluster ID...

Figure 5.3 Free energy profiles calculated from the BE-META simulations of chain B of insulin. The position of the lowest energy state of the neutral replica is marked by a blue square on each free energy curve. A typical lowest energy structure is also shown...

Figure 5.4 Representative low free energy cluster of the three basins determined using $p = 1.08$. Molten-globule 1 Most stable cluster of the basin, where all charged residues interact with the N- and C-termini of the protein, giving rise to strong electrostatic contacts...
Figure 5.5 Molten-globule 2 a) Most stable cluster of the basin, where Phe25 (green) is buried in a hydrophobic pocket, while Phe24 (yellow) is partially exposed to the solvent. b) This structure is only 0.9 kcal/mol from the most stable cluster and shows the presence of a hydrophobic core in which Phe24 (yellow) is present. Phe25 (green) is partially solvent exposed in this basin. c) In this cluster the Tyr16 (orange) residue and partially Phe24 (yellow) are buried in a hydrophobic pocket. The hydrophilic part of Tyr16 is directed toward the Arg11, Gln21, and the terminal salt bridges. 110

Figure 5.6 Native-like a) Lowest free energy cluster in the folded state basin. b) Cluster present between the native-like and molten-globule 2 basin having the N-terminal α-helix unfolded. 110

Figure 5.7 Occupancy (%) of the three basins found for $p = 1.08$ as a function of temperature. A simple linear extrapolation [65] from the calculated cluster enthalpy and entropy at 300 K has been used to obtain the basin occupancy versus temperature. 112

Figure 5.8 Metastable sets (cluster basins) detected by MCL using $p = 1.12$, are shown as coloured contours, with the coloured sphere corresponding to the lowest free energy cluster of each basin. The respective structures are presented above with the same colour code and ordered based on their free energy (1 has the lowest, 8 has the highest free energy). 113

Figure 5.9 The lowest energy structure (blue) from the folded state basin superimposed with the X-ray crystallographic structure of chain B, PDB code 1ZNI (pink). 116

Figure 5.10 Schematic representation of the basin dynamics. Transitions between the basins are represented by an arrow along with the corresponding transition time. Transitions that occur on a time scale longer than 1 µs are represented as dashed arrows. 118

Figure 6.1 Structure representation of apoC-II(60-70) monomeric peptide showing mutated residues and simulation times. 126

Figure 6.2 Secondary structure evolution plot for each mutation of apoC-II(60-70) peptide; a) Wild-type; b) oxi-Met; c) Met60Val; d) Met60Gln. The secondary structure colour codes: magenta – α helix, red – π helix, cyan – turn, white – coil, yellow – extended conformation and green – hydrogen bridge. 129
Figure 6.3 The number of clusters identified using an RMSD cutoff of 0.15 nm for each individual simulation is shown in the left of the figure. Structural representation of the two most populated clusters and their population are presented on the right.................................................................131

Figure 6.4 Radius of gyration (nm) of each simulation, represented by a moving average of 5 ns; a) Wild-type; b) oxi-Met; c) Met60Val; d) Met60Gln. ..........................134

Figure 6.5 Contact maps representing the average contact distance between all atoms of each residue for the structures contained in the most populated cluster of apoC-II(60-70) peptide; a) Wild-type; b) oxi-Met; c) Met60Val; and d) Met60Gln. A contact is defined by a distance of < 0.5 nm. .............................136

Figure 6.6 Histograms of Tyr63 ring orientation with respect to Phe67 (x-axis), obtained from total simulated trajectories for oxidised Methionine [285] (red), Met60Val (green), Met60Gln (blue) and wild-type (black) peptides. Y-axis indicates the percentage of structures for a specific orientation angle range exhibited during the trajectory. Forty bins were used over the range 0 – 180°. Each data point represents the percentage (%) of structures with orientation angles between x and x + 4.5°. Angle < 90° indicates both rings are on one side of the hairpin, while > 90° indicates rings exist on opposite sides. Structures illustrating the relative ring orientations are represented in insets. ...................................................................................................................138

Figure 6.7 Average values calculated per residue over the structures contained in the most populated cluster of the oxidised Methionine (red), Met60Val (green), Met60Gln (blue) and wild-type (black) simulation. a) Solvent accessible surface area (nm²); b) The average number of water molecules within a 3 Å shell of each residue........................................................................................................141

Figure 7.1 Structural representation of D5PC lipid molecule, where the carbon atoms are represented in cyan, hydrogen in red and the phosphate in green. The apoC-II(60-70) starting conformations, coil and helix are also shown together with the residue sequence of the peptide. .........................................................148

Figure 7.2 Typical radial distribution functions of the lipid head/s and tail/s with respect to hydrophobic and hydrophilic residues of the apoC-II(60-70) peptide. The RDFs for the 4lh system are shown as an example, while for all the other systems the RDF plots are presented in the Appendix of the thesis.................151
Figure 7.3 Schematic of the micellar-like formation of lipids surrounding the peptide. The hydrophobic residues are represented in blue, hydrophilic in red and the lipid in black. ..................................................................................................................152

Figure 7.4 Histograms representing the contacts formed between the lipid head and tail with individual residues of apoC-II(60-70) for systems Ilc and 6lh. Percentages of the times in contact are shown in the red/blue histograms, while the percentages of total number of contacts are shown in the grey/green histograms of each system. ..................................................................................................................154

Figure 7.5 Average solvent accessible surface area (SASA) per residue and corresponding error ..................................................................................................................156

Figure 7.6 Histogram representation of the % of lipids that were in contact with the peptide for every simulation. The bars of the histogram plots are coloured as described in the Computational Details section of this chapter. .................................159

Figure 7.7 Water radial distribution functions (RDF) with respect to apoC-II(60-70) peptide at various lipid concentrations; a) Water RDF profiles of the simulations starting from the coil conformation; b) Water RDF profiles of the simulations starting from the helix conformation. The region of the first RDF peak is shown as an inset. The RDF profiles are coloured as described in the Computational Details of this chapter. .................................................................161

Figure 7.8 Structural models of several lipid-peptide complexes. The surface of the peptide is coloured yellow, whereas the surface of the lipids is coloured red. .......164

Figure 7.9 Secondary structure evolution of apoC-II(60-70) at various lipid concentrations. Each simulation is labelled as described in the Computational Details section of this chapter. .................................................................165

Figure 7.10 Potential of mean force (PMF) profiles as a function of centre of mass separation (COM) of the N- and C-termini of apoC-II(60-70) peptide at various lipid concentrations. Magnified representations of selected regions are shown as an inset of the figure. Each PMF profile is coloured and labelled as described in the Computational Details section of this chapter. ......166

Figure 7.11 Average solvent accessible surface area (SASA) per residue. The calculations were performed on the data collected using umbrella sampling. Each plot is labelled as described in the Computational Details section of this chapter. ..................................................................................................................169
Figure 7.12 Water radial distribution functions (RDF) with respect to apoC-II(60-70) peptide at various lipid concentrations calculated from the simulations performed using umbrella sampling at the lowest free energy region. Magnified representation of the region of the first peak is shown as inset. The RDF profiles are coloured as described in the Computational Details of this chapter.

Figure 8.1 Illustration of the starting arrangement and size of apoC-II(60-70) oligomeric composites.

Figure 8.2 Root mean square deviation (RMSD) plots of the trajectories of dimeric apoC-II(60-70) oligomers. The structures of the most populated clusters of every system and their population are also presented.

Figure 8.3 Contact maps representing the average contact distances between the backbone and side-chain atoms between each residue of each dimeric system of apoC-II(60-70). The residues of monomer 1 are shown on the horizontal axis, whereas the residues of monomer 2 are shown on the vertical axis.

Figure 8.4 Root mean square deviation (RMSD) plots of the trajectories of trimeric apoC-II(60-70) oligomers. The structures of the most populated clusters of every system and their population are also presented.

Figure 8.5 Root mean square deviation (RMSD) plots of the trajectories of tetrameric apoC-II(60-70) oligomers. The structures of the most populated clusters of every system and their population are also presented.

Figure 8.6 Root mean square deviation (RMSD) plots of wild-type and mutated, four stranded anti-parallel oligomer of apoC-II(60-70).

Figure 8.7 Structures representing the typical conformations observed during the simulations of mutated tetrameric apoC-II(60-70) oligomers; a) 1-4-APC-oxiMet (red), b) 1-4-APC-M60V (green) and c) 1-4-APC-M60Q (blue). In some structures the aromatic side-chains are also shown, as well as the surface of the hydrophobic residues is represented as a green mesh.

Figure 8.8 Free energy profiles as a function of the center-of-mass (COM) separation of the wild-type and mutated (oxiMet60, Met60Val and Met60Gln) apoC-II(60-70) dimers with a) Charged and b) Neutral termini.
Abstract

In this thesis computer modelling studies were conducted to investigate protein behavior in various environments causing their folding, unfolding and aggregation. An introduction to the principles of protein structure and function, along with a concise literature review on some of the latest discoveries in the area of protein folding and aggregation are presented in Chapter 1.

Classical Molecular Dynamics techniques and their derivative methods such as umbrella sampling and bias-exchange metadynamics (BE-META) were employed and are described in Chapter 2. Applications related to two important proteins – insulin and apolipoprotein C-II (ApoC-II) are presented. The current knowledge of the structure and behaviour of these proteins is discussed in Chapter 3.

The use of atomistic simulation methodologies based on empirical force fields has enhanced our understanding of many physical processes governing protein structure and dynamics. However, the force fields used in classical modelling studies are often designed for a particular class of proteins and rely on continuous improvement and validation by comparison of simulations with experimental data. In Chapter 4 a comprehensive comparison of five popular force fields for simulation of insulin is presented. The effect of each force field on the conformational evolution and structural properties of the protein is analysed in detail and compared with available experimental data.

A fundamental phenomenon in nature is the ability of proteins to fold ab initio to their functional native conformation, also known as their biologically active state. Due to the heterogeneity and dimensionality of the systems involved, it is necessary to employ methodologies capable of accelerating rare events, specifically, configurational changes that involve the crossing of large free energy barriers. In Chapter 5, using the recently developed method BE-META the structural transitions and possible folding pathways of insulin were identified.

Another interesting phenomenon is the misfolding of proteins causing their aggregation, that may lead to formation of either amorphous compounds or structures of elongated-unbranched morphology known as amyloid fibrils. The deposition of amyloid fibrils in the human body may cause many debilitating diseases such as Alzheimer’s and
variant Creutzfeldt-Jakob diseases, thus making this field of research important and urgent. Due to the insoluble and non-crystalline nature of amyloid fibrils, experimental techniques are unable to elucidate the molecular mechanisms of fibril formation, in particular the initial stages of self-association. Thus computational methods are suitable for the investigation of these early fibril forming events and can give atomistic details of the initial peptide aggregation mechanisms. The human plasma protein apoC-II serves important roles in lipid transport, and it has been shown to form amyloid-like aggregates in solution. Recently, it has been demonstrated experimentally that oxidation of Met60 in the region of apoC-II(60-70) results in inhibition of fibril formation. Computational studies were performed to investigate the effect of mutations, such as Met oxidation and the residue substitutions to hydrophobic Val and hydrophilic Gln, on dynamics of apoC-II(60-70) peptide. The conformation features relevant to the amyloidogenic propensities of the peptide were identified and presented in Chapter 6.

The involvement of lipids at the various stages of development of amyloid diseases is becoming more evident in recent research efforts. In particular, micellar and sub-micellar concentrations have been shown to have different effect on fibril growth and kinetics of native apoC-II and derived peptides. In Chapter 7, investigation on the influences of phospholipids at various concentrations on the structure of apoC-II(60-70) using MD and umbrella sampling methods was performed. The molecular mechanisms of lipid effects on the peptide conformation and dynamics were identified.

In Chapter 8 preliminary results on the structural stability of pre-formed oligomeric composites of apoC-II(60-70) peptide of different sizes (dimer, trimer and tetramer) and arrangements (parallel and anti-parallel) were also presented. The most stable oligomer formation was a tetramer with the \( \beta \)-strands arranged in an anti-parallel conformation. The effects of mutation (oxidised Met, Met60Val and Met60Gln) on the most stable cluster were also investigated.

To conclude, several ideas for continuation of research in the protein folding and aggregation field are discussed in the Future Work section of this thesis.
Chapter 1

1. Computational studies of protein folding and aggregation

1.1 Overview

In this chapter, an overview of the current knowledge of the structure and dynamics of proteins, specifically their folding and aggregation is presented. A brief introduction on protein structure is presented in Section 1.2. A detailed literature review of the research efforts made in the field of protein folding and aggregation is presented in Sections 1.3 and 1.4, respectively. Finally, the project aims are described in Section 1.5.
1.2 Protein structure

Proteins are important biological macromolecules and an essential part of living organisms with many important roles. Proteins such as enzymes help catalyse biochemical reactions that are vital to metabolism. Protein hormones are involved in cell signalling, stimulation or inhibition of growth and immune responses, as well as regulating metabolism and controlling the reproductive cycle, while antibodies are used by the immune system to identify and neutralise foreign objects, such as bacteria and viruses.

A protein is essentially a polymer consisting of a long chain of amino acid residues, joined together via peptide bonds. There are 20 different amino acids that occur naturally in proteins. Each amino acid is composed of a central carbon atom (Cα) which is attached to a hydrogen atom (H), an amino group (NH₂), and a carboxyl group (COOH). Individual amino acids are distinguished from one another by a unique side-chain (R) attached to the Cα through its fourth valence, see Figure 1.1.

![Figure 1.1](image-url) The general structure of an amino acid, with the amino group on the left, the carboxyl group on the right and the unique side-chain (R) bonded to the Cα atom.

Amino acids are joined during protein synthesis by the formation of peptide bonds, by linking the carboxyl group of one amino acid to the amino group of another. During this process a water molecule is lost, as illustrated in Figure 1.2. A polypeptide chain consists of a regularly repeating part, called the main-chain or backbone (NH–CαH–C'=O), and a variable R-group, comprised of distinctive side-chains. It is the chemical nature of the side-chain components of individual residues that determines the property of the protein. The properties are usually divided into three different groups.
The first group comprises amino acids with strictly hydrophobic side-chains: Ala (A), Val (V), Leu (L), Ile (I), Phe (F), Pro (P), and Met (M). The four charged residues, Asp (D), Glu (E), Lys (K), and Arg (R), form the second class. The third class comprises those with polar side-chains: Ser (S), Thr (T), Cys (C), Asn (N), Gln (Q), His (H), Tyr (Y), and Trp (W). The amino acid Gly (G) is the simplest of the 20 amino acids; it has only a hydrogen atom as a side-chain, therefore has special properties and is usually considered to be a member of the hydrophobic amino acid group. Cysteine residues have a unique property in their ability to form disulfide bridges with another Cys side-chain in an oxidative environment. As the two cystines are covalently bonded, the disulfide bond is very strong, and is essential for the folding, structure and function of a protein. In some proteins, these bridges hold together different polypeptide chains, such as the A and B chain of insulin, which are linked by two separate disulfide bonds (an example of this type of bond is shown in Figure 3.1, Chapter 3).

All amino acids by glycine are chiral molecules. That is, they exist in two optically active, asymmetric forms, called enantiomers that are mirror images of each other. They have been designated as having a D- and L-configuration. The amino acids found in proteins...
CHAPTER 1. Computational studies of protein folding and aggregation

almost always possess the L-configuration. Some D-amino acids are found in microorganisms, particularly in the cell walls of bacteria.

Proteins are assembled from amino acids using the information encoded in our DNA (deoxyribonucleic acid). However, the process of protein formation or synthesis is conducted over several important steps. Initially, the DNA transcribes a messenger RNA (mRNA) through transcription, where a small copy of a gene is released to the cytoplasm. Translation of the mRNA occurs in the ribosomes which are located in the cytoplasm. During translation, the mRNA is decoded to produce a specific polypeptide chain according to the rules specified by the genetic code.Shortly after or during synthesis, the residues in the proteins are often chemically modified by post-translational modification, where the physical and chemical properties of the protein are altered, resulting in changes in the fold, stability and function of the protein. The importance of correct folding and folding behaviour in proteins is a key topic of this thesis. A detailed review on the current knowledge of protein folding is presented in the following sections.

The functional properties of proteins depend on their three-dimensional structure, which for convenience is represented in several structural levels. A schematic representation of the different levels of protein structure is shown in Figure 1.3. The amino acid sequence of a protein’s polypeptide chain is called its primary structure. Different regions of the sequence form local regular secondary structures, such as α-helices or β-strands. The three essential torsion angles: φ, ψ, and ω determine the conformation of the protein backbone. By convention, the angle of rotation around the N–Ca bond is called the phi (φ) angle, while the rotation around the Ca–C bond is called psi (ψ). The rotation around the C–N bond is called omega (ω), and this angle has a value close to 180° (except for cis-prolines, which have ω-angle close to 0°). The secondary structure is formed when consecutive residues have similar dihedral angle (φ, ψ) values that lie within known regions on the Ramachandran map [1]. There are three sterically allowed regions on the Ramachandran plot, illustrated in Figure 1.4. These regions correspond to secondary structures with well-defined dihedral angles: the right-handed α-helix, the left-handed α-helix and the β-strand. The α-helix is a classical element of protein structure, and is easily recognised by its spiral-look. This structural motif is stabilised by the formation of parallel hydrogen bonds between the backbone atoms of consecutive residues in a protein. There are three types of commonly occurring helices: α-helix, 3_{10}-helix, and π-helices, all varying according to the hydrogen bond coordination of their backbone atoms, making the chain either more loosely or more tightly coiled.
The amino acid sequence of a protein is called its **primary** structure. Different regions of the sequence form local regular **secondary** structures, such as α-helices or β-strands. The **tertiary** structure is formed by packing such structural elements into a globular unit or domain. The final conformation of a protein may contain several polypeptide chains arranged in a **quaternary** structure.
The less energetically favourable helices are the $3_{10}$ and $\pi$-helices, which have bonds between residues $i$ and $(i + 3)$, $i$ and $(i + 5)$, respectively. An ideal $\alpha$-helix has bonds between residues $i$ and $(i + 4)$ and has dihedral angles of (-57.8°, -47.0°), corresponding to the bottom left quadrant of the Ramachandran plot.

The second major structural element found in globular proteins is a $\beta$-strand. The $\beta$-strand has a predisposition to self-associate and to form $\beta$-sheets, which are composed of several $\beta$-strands bound together by hydrogen bonds between the backbone atoms. The $\beta$-strands have an extended conformation, with average dihedral angles of (-120°, 120°), which belong in the broad region in the upper left quadrant of the Ramachandran plot. These strands interact in two ways to form pleated $\beta$-sheets: parallel and anti-parallel, as demonstrated in Figure 1.5. The $\beta$-sheet structures are associated with the formation of amyloid fibrils, which are known to be the cause of many debilitating diseases.

**Figure 1.4** Ramachandran diagram showing the three sterically allowed regions for common secondary structures of a protein. Lighter (yellow/green) colours represent the more favourable combinations of $\psi$-$\phi$ angle values.
Most protein structures are built up from a combination of secondary structure elements, such as helices and β-strands that are connected by turns and loop regions of various lengths and shapes. These structural elements are essential for protein folding and the formation of compact tertiary structures, as well as enabling flexibility in the protein, which is often related to its activity. Loop regions that connect two adjacent anti-parallel β-strands are called β-hairpins, shown in Figure 1.6. The formation of tertiary structure is mostly driven by hydrophobic interactions between the secondary structure elements which minimise the solvent accessible surface area and increase the protein’s conformational stability [2].

Figure 1.5 Illustration of the hydrogen bonding patterns represented by dotted red lines in a) parallel; and b) anti-parallel β-sheets.
The arrangement of multiple folded protein molecules in a multi-subunit complex is referred to as a quaternary structure. An example is the insulin dimer, consisting of two distinct monomer structures bonded together by disulfide bonds, which further combine into a hexamer in the presence of metal ions, such as zinc. An illustration of the quaternary hexamer structure of insulin is shown in Figure 1.3.

The correct three dimensional structure or native state is essential for protein function. Failure to fold into the native shape usually produces proteins with different properties that sometimes can be toxic, such as the prion protein which is associated with variant Creutzfeldt-Jakob disease. Several neurodegenerative and other diseases result from the accumulation of misfolded proteins. A literature review on significant discoveries in the area of protein folding and aggregation is presented in the following sections.
1.3 Protein folding

One of the most fundamental phenomena in nature is the capability of proteins to fold \textit{de novo} to their native conformation, also known as their biologically functional state. A major milestone in protein science was the hypothesis of Christian Anfinsen and co-workers [3], who proposed that the native structure of a protein is the thermodynamically stable conformation, which depends only on the amino acid sequence and the conditions of the solution, not on the kinetic folding route. To reach this stable native structure, the physical process of protein folding occurs, where the polypeptide chain folds into its three dimensional state. However, a random search to sample all possible conformations to reach the native state through the vast conformational space available to a protein is unfeasible. For example, a 101 amino acid sized protein would have to sample $3^{100}$ configurations before finding its native state, which could take years to achieve. However, it has been established that proteins fold within a timeframe of microseconds to seconds. This became known as the Levinthal’s paradox [4]. In 1968, Cyrus Levinthal concluded that random searches are not an effective way to find the correct folded state of a protein [5]. Since then, computational and theoretical advances have aimed to shed some light on the protein folding problem, and have complemented experiments by elucidating some of the folding mechanisms at atomic detail. However, to this day the folding process of a linear polypeptide strand to its three-dimensional, biologically active conformation is poorly understood and theoretical prediction of the folding pathways remains a challenge.

Some of the theoretical and computational methods used to elucidate the protein folding problem are presented below. The proposed mechanisms of folding in proteins are also discussed. Some of the more significant computational discoveries made in protein folding are summarised.

1.3.1 Studies of protein folding mechanisms

Initially, the models used for protein folding studies were very simple and could not be used to observe atomic-level events of the folding process. Specifically, protein folding was traditionally studied by residue-level models [6]. With these methods the accuracy of results is sacrificed by the reduced level of representation. One example of such a method is the Gō-model which makes use of a potential function that is based on the knowledge of the native
CHAPTER 1. Computational studies of protein folding and aggregation

structure of the protein. As a rule, the potential is the sum of two-body terms where if a native contact is formed the result is -1 and 0 otherwise. Specifically, the Gō-model considers only the interactions between residues (beads on the lattice) that are present in the native state, thus only native contacts are taken into account. This interaction model is commonly used in thermodynamic and kinetic studies of protein folding to model amino acid interactions [7, 8].

Over time a more physical model, typically based on polymer-like lattices, was developed and applied to study the static and dynamic properties of protein conformational space [9, 10]. The application of computer simulations using simplified lattice and off-lattice models greatly enhanced our understanding of various aspects of the protein folding problem. The simplicity and computational efficiency of these models made it possible to simulate thousands of folding and unfolding events. Although these models do not represent the full complexity of real proteins, they capture the core aspect of the physical protein folding problem, having the capability of finding the lowest energy state without an exhaustive conformational search. Using these methods detailed statistical description of the folding process in protein models could be obtained, as reviewed by Mirny and Shakhnovich [6].

One of the early proposed mechanisms for protein folding was the nucleation-condensation model [11-13], which was tested in the context of lattice models [13]. The nucleation-condensation model postulated that a small number of residues, referred to as a folding nucleus, need to form their native contacts in order for the folding reaction to proceed to the native state. Based on this model, and its analogy to first-order phase transitions, the concepts of nucleation and free energy landscape have prompted much of the recent progress in understanding the process of protein folding.

Several other models emerged in the attempt to elucidate the protein folding mechanism. One such model is the pathway model, which implies the existence of only one route for a protein to fold or unfold. Based on Levinthal’s suggestion of the existence of a folding pathway which allows proteins to fold in a realistic timeframe, this model satisfies the time constraint. In the pathway model the existence of intermediate states (I) in relation to the folded (N) and unfolded (U) states, are classified as on-pathway [14]:

\[ U \leftrightarrow I \leftrightarrow N \]

and off-pathway [14]:

\[ U \leftrightarrow N \]

\[ I \]

12
CHAPTER 1. Computational studies of protein folding and aggregation

with off-pathway intermediates believed to take no part in the actual folding process. It soon became evident that there are many different pathways through which a protein can fold, each with its own set of transition states. Proteins are generally thought to exhibit funnelled energy landscapes which allow proteins to fold to their native states through a stochastic process in which the free energy decreases spontaneously. The unfolded state, transition state, native state and possible intermediates correspond to local minima or saddle points on the free energy landscape. The concept of “folding funnel” energy landscape was introduced by Leopold et al. [15]. The team investigated the folding of two different 27-unit chains using the lattice modes, and found that one was able to fold and the other was not. Leopold suggested that “convergent kinetic pathways or folding funnels guide folding to a unique, stable native conformation”, thus kinetic trapping prevented the second structure of folding. Four generalised types of funnel energy surfaces that represent possible folding mechanisms of proteins are illustrated in Figure 1.7 (adopted from reference [16]). The figure shows the different energy landscapes: fast folding (simple funnel), kinetic trapping (with one well or rugged with many wells), and slow random searching (golf course).

![Figure 1.7 Cartoon representation of the different energy landscapes from a denatured conformation to the native conformation N, adopted from reference [16].](image)

- (a) A smooth energy landscape for fast folding protein;
- (b) A moat landscape, where folding must go through an obligatory intermediate;
- (c) A rugged energy landscape with kinetic traps;
- (d) A golf course energy landscape in which folding is dominated by diffusional conformational search.

Some of the initial computational investigations of protein folding in atomic detail consisted of energy minimisation methods which were applied to protein structures [17, 18]. This method was followed by molecular dynamics [19, 20] which is used to this day. All-atom protein models with explicit and implicit solvents enabled the investigation of folding thermodynamics and unfolding dynamics of small proteins [21-23]. However, due to the
CHAPTER 1. Computational studies of protein folding and aggregation

complexity and large dimensionality of the protein conformational space, all-atom MD simulations have severe limitations on the time and length scales that can be studied (discussed in Chapter 2).

Novel simulation techniques have been developed that improve the conformational sampling efficiency, including biased sampling of the free energy surface and non-equilibrium unfolding simulations. Generalised ensemble sampling techniques, that involve parallel simulation of molecular systems coupled with a Monte Carlo protocol [24] have been successfully applied to investigate protein folding. Using the Replica Exchange Monte Carlo (REMC) method [25], classical Replica Exchange Molecular Dynamics (REMD) [26] and multiplexed REMD [27], the folding dynamics of small proteins were obtained. The methodologies behind some of the more widely used enhanced sampling methods, such as Umbrella Sampling, REMD and the recently developed method Bias-Exchange Metadynamics are discussed in Chapter 2. Excellent reviews on the need for improved conformational sampling and the latest associated methods are presented by the group of Shea and van Gunsteren [28, 29].

Alternatively, to improve conformational sampling a large number of short simulations can also be performed. The world-wide parallel computing network, such as Folding@Home, utilises many processors in a highly heterogenous and loosely coupled distributed computing paradigm to increase the computational efficiency in running MD. Using this approach, Pande and co-workers were able to accumulate hundreds of microseconds of atomistic MD. The folding mechanism and folding rate of several fast-folding proteins and polymers were determined with good accuracy to experimental data [30].

Multi-scale modelling approaches have also been used to combine efficient conformational sampling of coarse-grained models and accuracy of all-atom models to study protein folding pathways. In this approach, iterative simulations and inter-conversion between high and low-resolution protein models are performed. Feig et al. developed a multi-scale modelling tool set, called MMTSB [31]. This method integrates a simplified protein model (lattice-based low resolution conformational sampling) with the Monte Carlo simulation engine, MONSSTER [32], and for all-atom simulations incorporates the MD packages AMBER [33] or CHARMM [34]. Ding et al. reconstructed the transition state ensemble of the src-SH3 protein domain through multi-scale simulations [35].

The protein folding studies can also be facilitated by sampling protein conformations near the native state. Several native-state sampling algorithms, such as normal mode analysis
(NMA) and the structure-based algorithm COREX [36, 37] have been successfully utilised to study plasticity [38], cooperative interactions [39] and allostery [40] in proteins. Native-state ensemble techniques take into account protein flexibility, which is important in biological activity and crucial in structure based drug design.

In the last five years several tools for performing web-based analysis of protein folding dynamics have been developed. The Fold-Rate server (http://psfs.cbrc.jp/fold-rate/) [41] predicts rates of protein folding using the amino-acid sequence. The Parasol folding server (http://parasol.tamu.edu(6,7),(993,990) groups/amatogroup/goldingserver) [42] predicts protein folding pathways using “probabilistic roadmaps”-based motion planning techniques. The iFold server (http://ifold.dokhlab.org) allows discrete molecular dynamics (DMD) simulation of protein dynamics using simplified two-bead per residue protein models [43]. DMD simulations solve ballistic equations of motion with square-well approximation to the inter-particle interaction potentials. DMD approaches [44-46] with simplified structural models of proteins have been extensively used for investigating the general principles of protein folding and unfolding [47-51].

Advances in experimental techniques have also been made, such as protein engineering, nuclear magnetic resonance (NMR), mass spectrometry, hydrogen exchange, fluorescence resonance energy transfer (FRET) and atomic force microscopy (AFM). These techniques have made it possible to obtain detailed information about the different conformations occurring in the folding process [52]. At the same time, computational methods have been developed to better interpret experimental data by using simulations to obtain structural information about the states which are populated during the folding process [53]. The synergy between experiment and theoretical methods is increasing because the timescales and resolutions at which they operate are merging.

### 1.3.2 Protein folding kinetics

Extensive studies using MD simulations have been conducted with the aim of predicting the folding rates of some protein models. A challenge of this approach was whether sufficient sampling and simulation time scale can be obtained where the folding transitions can be observed. Early studies pioneered by Caflisch et al. employed continuum solvent with low viscosity to observe multiple folding transitions [54-56]. However, there is a nonlinear relationship between the folding time and viscosity [57]. To circumvent this problem, Pande
et al. used coupled ensemble dynamics to simulate the folding of a β-hairpin from protein G using a continuum solvent model and the united-atom force field OPLS with water-like viscosity. In this study the folding rate of the β-hairpin was calculated in close agreement with experimental results. One of the more significant initial efforts in protein folding studies was that of Duan and Killman [21]. Using parallel computers with increased efficiency, an extensive 1 µs folding simulation on the villin headpiece was performed. From their results they determined two distinct folding pathways of the villin protein.

Rate predictions using MD simulations with explicit solvent have been performed to gain additional insight into the folding kinetics in the presence of water. Pande et al. yielded experimentally consistent rates when they investigated helix to coil transitions [58] and folding rates of the artificial mini-protein BBA5 [59]. In a later study the distributed computing implementation (Folding@Home) was used to produce tens of thousands of 5-20 ns trajectories, totalling 700 µs, which allowed the folding rate of this small protein to be determined.

Protein folding kinetics have been investigated with other methods such as DMD. Jang et al. used DMD and simplified protein models with Gō interactions to probe protein folding kinetics [48]. Protein folding kinetics studies using DMD simulations are reviewed in detail by Dokholyan et al. [60].

Folding rate predictions using classical MD principles have been limited to small two-state folding peptides, such as those described above. For larger proteins it is difficult to computationally study folding kinetics due to the inability of classical methods to effectively sample large conformational space. Chapter 2 discusses some of the methods available today capable of effective and efficient sampling of the protein conformational space.

### 1.3.3 Transition / intermediate states in protein folding

Experimentally, structural investigation of the transition state ensemble (TSE) is extremely challenging, since it is an unstable state which can be very difficult to detect. However, computational methods can be applied to identify transition state conformations using path sampling, unfolding simulations or by projection into one of two reaction coordinates (Figure 1.8).
The intuitive implementation of a low-dimensional energy landscape in explaining simple chemical reactions was adopted in similar formulation for protein folding investigations. However, unlike simple molecules, a protein has a high number of degrees of freedom which makes the analysis very difficult. To circumvent this problem, several groups have proposed dimensional reduction by projecting the multi-dimensional energy landscape into few relevant coordinates [9, 61]. Others directly tackled the transition state issue by developing rigorous path sampling methods, such as the Transition Path sampling technique as described by Bolhuis et al. [62]. These methods allow the sampling of rare events without prior knowledge of the reaction coordinates, mechanisms or transitions states. However, these methods can be very computationally expensive. More recently, a method called Bias-Exchange Metadynamics (BE-META) was developed which has proven to effectively sample the conformational space of protein using several reaction coordinates with great computational efficiency [63]. The BE-META technique was successfully applied in several studies of protein folding [64-66], protein-protein interactions and enzyme reactions [67].

Characteristics of the transition state ensemble have been mainly investigated by \( \Phi \)-analysis [68], which involves measuring the folding kinetics and equilibrium thermodynamics of mutants containing amino acid substitutions throughout a protein. Experiments suggest that a protein may be kinetically trapped \textit{en route} to the native state [69, 70], however, the mechanism by which proteins avoid these kinetic traps is still unclear. By using a G\( \ddot{o} \)-model scaled to include sequence-specific interactions, Khare \textit{et al.} found that the residues which contribute most to Cu, Zn SD1 stability also function as “gatekeepers” that avoid kinetic traps and protein misfolding [71]. These type of “gatekeeper” residues were also identified in later
computational studies of the ribosome protein S6, where Stoycheva et al. [72] and Matysiak and Clementi [73] showed that the mutations of these residues can alter the folding landscape of S6 and shift the balance between its folding and aggregation.

As previously described proteins sample ensembles of heterogeneous conformations in solution. In recent years, it has been found that even for the small two-state proteins (~100 amino acids or less) there exist partially unfolded intermediates on the folding pathways. These intermediates are difficult to detect in kinetic folding experiments [74, 75]. However, growing evidence has indicated that the intermediate states formed during protein folding and unfolding may have significant roles in protein function by exposing post-translation modifications or ligand binding sites. The intermediate states are usually weakly-populated (thermodynamic intermediates) or short-lived (kinetic intermediates). Recent computational and experimental studies have greatly facilitated the unprecedented structural characterisation of rare intermediates and suggested a functional role for these conformations.

Gsponer et al. [76] and Dixon et al. [77] have developed computational approaches to incorporate hydrogen exchange protection factors from NMR experiments as constraints into MD simulations to detect and/or characterise the conformations of intermediate states. Using this new approach, the folding intermediates of the bacterial immunity protein Ig7 were defined [76]. Additionally, Dixon et al. were able to characterise a folding intermediate of the focal adhesion targeting (FAT) domain of focal adhesion kinase (FAK), which plays critical roles in cell proliferation and migration [77]. Using steered molecular dynamics (SMD) the unfolding intermediate of Ig domain I27 from the titin protein which is responsible for muscle elasticity, and domain 10 of type III fibronectin module were predicted [78, 79].

1.3.4 Unfolded or denatured states

Understanding the structural and dynamic properties of denatured proteins is crucial for understanding the protein folding [80, 81] and misfolding [82] events. For example, the computational determination of a protein’s thermodynamic stability requires an accurate approximation of the denatured state as the reference state. Additionally, understanding the structural properties of unfolded proteins may shed light on the early events of protein folding and aggregation [83].

It has been postulated that the denatured state of proteins is composed of an ensemble of featureless random-coil like conformations, however the possibility that denatured proteins
can have residual native-like structures has not been excluded. Increasing experimental evidence shows that under denaturating conditions, the protein can still exhibit conformational bias and retain a certain amount of native structure [84, 85]. Several theoretical and computational studies have addressed the role of specific interactions in conformational biasing toward the native state whilst occupying the denatured state. Using a simple force field with only steric and hydrogen bond interactions, Pappu et al. [86] demonstrated that denatured protein states have a strong preference for native structure. It was suggested in separate studies [86-88] that the conformational bias of native structures in the denatured state is a possible explanation of the Levinthal’s paradox. Ding et al. developed a computational method to model denatured proteins using a structure-based potential of the Gō-model [89]. This study suggested that denatured proteins follow the random coil scaling sizes obtained experimentally and retain residual secondary structures akin to those observed in native protein states.

The observation of residual native secondary structures in thermally-denatured protein states is consistent with a “guided-folding” scenario [90], where the rate-limiting process is the packing of preformed secondary structures into the correct fold. In contrast, a random coil model of the denatured state without residual native-like structures implies that a protein has to overcome a high entropic barrier to form both the secondary and tertiary structure upon folding. The existence of persistent native-like secondary structures in the denatured state is responsible for the success of protein structure prediction methods like ROSETTA, where small secondary structure segments derived from the protein data bank are used as a template to make de novo structure predictions [91]. Furthermore, for the past ten years CASP (Critical Assessment of Structure Prediction) has monitored the state of the art in modelling protein structure from sequence. CASP provides users of structure prediction servers with an opportunity to assess the quality of various methods. An excellent review of the progress, bottlenecks and prognosis in protein structure prediction using CASP was presented by John Moult, and references within [92].

In summary, the field of protein folding has seen significant advances over the last few decades through the development of experimental and computational techniques. Following this discussion of some of the more recent discoveries and models proposed for the mechanisms of protein folding, it is evident that despite the significant progress made, there is a need for continued research into the ambiguities of the process and pathways of protein folding.
CHAPTER 1. Computational studies of protein folding and aggregation

1.4 Protein self-association and aggregation

While all the information needed for proteins to fold is encoded in their amino acid sequence, there are many more elements that play a part in vivo. In a crowded cellular environment, surrounded by interacting proteins, nascent polypeptides face a formidable challenge in finding the correct interactions that result in a folded and functional protein. Many become trapped in meta-stable intermediate structures which are usually recognised by proteasomal machinery and degraded or refolded by chaperones. Alternatively, they can also misfold and self-associate leading to formation of either amorphous compounds or structures of elongated-unbranched morphology, known as amyloid fibrils. An accumulation of these fibrils can result in a range of human diseases commonly referred to as amyloidosis. Some of the diseases associated with amyloid fibrils are Alzheimer’s, variant Creutzfeldt-Jakob disease, Parkinson’s, type II diabetes and many others.

1.4.1 Early history of amyloid diseases

The history of amyloidosis usually begins with a reference to Rudolf Virchow’s paper from 1854 who was first to use the term “amyloid” for structural deposits discovered in human tissue [93]. Virchow used a water solution of iodine in combination with hydrated sulphuric acid as a stain for cellulose in the human body [94], upon which he found that corpora amylacea in the ependyma and choroid plexus showed a typical cellulose reaction with iodine. Tissue testing was performed by Virchow and Meckel on samples of what we call today systemic amyloidosis, and found to have a similar reaction to iodine as had been observed with corpora amylacea [95]. These wax-like deposits were discovered to have degenerative effects on the spleen, liver and kidneys.

It was originally thought that amyloid was cellulose by nature. However, it was Friedreich and Kekulé who dissected the amyloid-rich segments from the spleen of a patient with amyloidosis [96]. In contrast to Virchow, they performed direct chemical analysis of the material extracted and reached the definitive conclusion that the main substance was protein.

Iodine staining was the initial method used to identify amyloid, as introduced by Virchow. This method was replaced by metachromatic stains like crystal violet. It was Bennhold [97] who introduced the most important histological stain marker for amyloid, Congo red. This dye has been used as a cotton colourant in the textile industry since 1884.
Congo red was observed to have strong affinity for amyloid deposits and could be used as a diagnostic method for amyloidosis. However, the significance of Congo red grew with the discovery made by Divry and Florkin in 1927 who noticed the enhanced birefringence (or double refraction) of amyloid deposits after staining with Congo red [98]. It was suggested that this property of amyloid depends on an ordered arrangement of the elongated Congo red molecules in the amyloid, and that amyloid was not amorphous as earlier described, but has an organised structure [99]. A standardised Congo red staining method was introduced by Puchtler et al. [100] and is still used today. Additional staining methods have been introduced such as Thioflavin T and S, which are also among the frequently used staining methods.

### 1.4.2 Protein aggregation related diseases

A list of known diseases that are associated with the formation of extracellular amyloid fibrils or intracellular inclusions with amyloid-like characteristics is given in Table 1.1. The proteins that are associated with each disease are listed, together with their chain length and known monomeric structure. The diseases can be broadly grouped into several categories. **Neurodegenerative conditions**, in which aggregation occurs in the brain. These include some of the more well known conditions such as Alzheimer’s and Parkinson’s diseases which are affecting a growing number of our ageing population. **Nonneuropathic localised amyloidosis** relates to a group of diseases where aggregation occurs in a single type of tissue other than the brain and **nonneuropathic systemic amyloidosis** relates to diseases in which aggregation occurs in multiple tissues.

Some of these conditions, such as Alzheimer’s and Parkinson’s diseases, are predominantly sporadic, although hereditary forms are well documented. Other conditions, such as the lysozyme and fibrinogen amyloidosis, arise from specific mutations and are hereditary. In addition to sporadic amyloidosis, which accounts for 85% of the amyloid diseases, and hereditary forms for 10%, spongiform encephalopathies can also be transmissible (5%) in humans as well as in other mammals.
## CHAPTER 1. Computational studies of protein folding and aggregation

<table>
<thead>
<tr>
<th>Disease</th>
<th>Aggregating protein or peptide</th>
<th>Number of residues(^a)</th>
<th>Native structure of protein or peptide(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Neurodegenerative diseases</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alzheimer’s disease(^c)</td>
<td>Amyloid β peptide</td>
<td>40 or 42(^d)</td>
<td>Natively unfolded</td>
</tr>
<tr>
<td>Spongiform encephalopathies(^c,e)</td>
<td>Prion protein or fragments thereof</td>
<td>253</td>
<td>Natively unfolded (1-120) and α-helical (121-230)</td>
</tr>
<tr>
<td>Parkinson’s disease(^c)</td>
<td>α-Synuclein</td>
<td>140</td>
<td>Natively unfolded</td>
</tr>
<tr>
<td>Dementia and Lewy bodies(^c)</td>
<td>α-Synuclein</td>
<td>140</td>
<td>Natively unfolded</td>
</tr>
<tr>
<td>Frontotemporal dementia with Parkinosis(^c)</td>
<td>Tau</td>
<td>352-441(^f)</td>
<td>Natively unfolded</td>
</tr>
<tr>
<td>Amyotrophic lateral sclerosis(^c)</td>
<td>Superoxide dismutase (^l)</td>
<td>153</td>
<td>All-β</td>
</tr>
<tr>
<td>Huntington’s disease(^d)</td>
<td>Huntingtin with polyQ expansion</td>
<td>3144(^g)</td>
<td>Largely natively unfolded</td>
</tr>
<tr>
<td>Spinocerebellar ataxias(^d)</td>
<td>Ataxins with polyQ expansion</td>
<td>816(^h)</td>
<td>All-β</td>
</tr>
<tr>
<td>Spinal and bulbar muscular atrophy(^d)</td>
<td>Androgen receptor with polyQ expansion</td>
<td>919(^g)</td>
<td>All-α</td>
</tr>
<tr>
<td>Familial British dementia(^d)</td>
<td>ABri</td>
<td>23</td>
<td>Natively unfolded</td>
</tr>
<tr>
<td>Familial Danish dementia(^d)</td>
<td>ADan</td>
<td>23</td>
<td>Natively unfolded</td>
</tr>
<tr>
<td><strong>Nonneuropathic systemic amyloidoses</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AL amyloidosis(^c)</td>
<td>Immunoglobulin light chains or fragments</td>
<td>~90(^f)</td>
<td>All-β</td>
</tr>
<tr>
<td>AA amyloidosis(^c)</td>
<td>Fragments of serum amyloid A protein</td>
<td>76-104(^f)</td>
<td>All-α</td>
</tr>
<tr>
<td>Familial Mediterranean fever(^c)</td>
<td>Wild-type transthyretin</td>
<td>127</td>
<td>All-β</td>
</tr>
<tr>
<td>Senile systemic amyloidosis(^c)</td>
<td>Mutants of transthyretin</td>
<td>127</td>
<td>All-β</td>
</tr>
<tr>
<td>Familial amyloidosic polyneuropathy(^d)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hemodialysis-related amyloidosis(^d)</td>
<td>β2-microglobulin</td>
<td>99</td>
<td>All-β</td>
</tr>
<tr>
<td>ApoA-I amyloidosis(^d)</td>
<td>N-terminal fragments of apolipoprotein A-I</td>
<td>80-93(^f)</td>
<td>Natively unfolded</td>
</tr>
<tr>
<td>ApoA-II amyloidosis(^d)</td>
<td>N-terminal fragments of apolipoprotein A-II</td>
<td>98(^e)</td>
<td>Unknown</td>
</tr>
<tr>
<td>Finnish hereditary amyloidosis(^d)</td>
<td>Fragments of gelsolin mutants</td>
<td>71</td>
<td>Natively unfolded</td>
</tr>
<tr>
<td>Lysozyme amyloidosis(^d)</td>
<td>Mutants of lysozyme</td>
<td>130</td>
<td>α+β, lysozyme fold</td>
</tr>
<tr>
<td>Fibrinogen amyloidosis(^d)</td>
<td>Variants of fibrinogen α-chain</td>
<td>27-81(^f)</td>
<td>Unknown</td>
</tr>
<tr>
<td><strong>Nonneuropathic localised diseases</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type II diabetes(^c)</td>
<td>Amylin, also called islet amyloid polypeptide (IAPP)</td>
<td>37</td>
<td>Natively unfolded</td>
</tr>
<tr>
<td>Medullary carcinoma of the thyroid(^d)</td>
<td>Calcitonin</td>
<td>32</td>
<td>Natively unfolded</td>
</tr>
<tr>
<td>Hereditary cerebral haemorrhage with amyloidosis(^d)</td>
<td>Mutants of amyloid β peptide</td>
<td>40 or 42(^d)</td>
<td>Natively unfolded</td>
</tr>
<tr>
<td>Injection-localised amyloidosis</td>
<td>Insulin</td>
<td>21+30(^h)</td>
<td>Mostly folded in α-helical conformation (Chapter 3)</td>
</tr>
<tr>
<td>Aortic medial amyloidosis(^c)</td>
<td>Medin</td>
<td>50(^e)</td>
<td>Unknown</td>
</tr>
<tr>
<td>Cataract(^c)</td>
<td>γ-Crystallins</td>
<td>Variable</td>
<td>All-β</td>
</tr>
<tr>
<td>Pulmonary alveolar proteinosis(^d)</td>
<td>Lung surfactant protein C</td>
<td>35</td>
<td>Unknown</td>
</tr>
<tr>
<td>Inclusion-body myositis(^c)</td>
<td>Amyloid β peptide</td>
<td>40 or 42(^g)</td>
<td>Natively unfolded</td>
</tr>
<tr>
<td>Cutaneous lichen amyloidosis(^c)</td>
<td>Keratins</td>
<td>Variable</td>
<td>Unknown</td>
</tr>
</tbody>
</table>
CHAPTER 1. Computational studies of protein folding and aggregation

---

Data refers to the number of residues of the processed polypeptide chains that deposit into aggregates, not of the precursor protein.

According to Structural Classification Of Proteins (SCOP), these are the structural class and fold of the native states of the processed peptides or proteins that deposit into aggregates prior to aggregation.

Predominantly sporadic, although in some cases hereditary forms associated with specific mutations are well documented.

Predominantly hereditary, although in some cases sporadic forms are documented.

Five percent of the cases are transmitted.

Fragments of various lengths are generated and have been reported to be present in ex vivo fibrils.

Lengths shown refer to the normal sequences with nonpathologic traits of polyQ.

Length shown refer is for ataxin-1.

The pathogenic mutation converts the stop codon into a Cys codon, extending the 77-residue protein by 21 additional residues.

Human insulin consists of two chains (A and B, with 21 and 30 residues, respectively) covalently linked by disulfide bridge.

Medin is the 245-295 fragment of human lactadherin.

---

Table 1.1 A list of diseases associated with the aggregation of proteins, adopted from the review article by Chiti and Dobson [101]. The name of the disease, the pathological protein, together with the size and native structure are specified.

1.4.3 Structure of amyloid fibrils

In 1935, Astbury and Dickinson placed a poached egg white in an X-ray beam and observed a diffraction pattern with perpendicular reflections at ~4.7 Å along the meridional direction and ~10 Å along the equatorial direction [102]. The pattern suggested that the protein chains of the egg white pack in an extended β-conformation, with the chains perpendicular to the long axis. This cross-β X-ray diffraction pattern was observed for the amyloid deposits in most diseased tissues. In 1959, Cohen and Cakins found that the amyloid, which is hyaline and structureless under light microscopy, in fact has a characteristic fibrillar ultrastructure when analysed with an electron microscope [103]. This finding was confirmed in several other studies [104-108] which gave further insight into the specific structural organisation of amyloid fibrils. Of particular interest was that although the morphology of the fibrils is similar, consisting of a cross-β structure with the β-strands perpendicular to the fibril axis, the proteins from which they are built are different.

Recently, significant progress has been made in determining the structures of amyloid fibrils in atomic detail using experimental techniques [109-111]. The fibril architecture of a number of peptides, including Aβ_{1-40} and Aβ_{1-42} have been proposed using constraints obtained from solid state NMR [109], and are also consistent with molecular dynamics simulations [112]. In addition, advances in micro-crystallographic techniques enabled the discovery of a high resolution structure of a peptide extracted from N-terminal segments of Sup35 and many others, reported by Eisenberg’s group [110]. Their studies have confirmed
that many peptides, which are unrelated by sequence, adopt the characteristic cross-\( \beta \) pattern in the fibril state. The crystal structure and characteristics, such as strand and sheet separations of the \( \text{A}\beta_{1-42} \) and Sup35 derived peptide, GNNQQNY, are shown in Figure 1.9.

\[ \text{A}\beta_{1-42} \]

\[ \text{GNNQQNY} \]

**Figure 1.9** Experimentally determined crystal structures of A\( \beta_{1-42} \) and Sup35 derived GNNQQNY peptides.

### 1.4.4 Mechanisms of fibril formation

Protein fibrillisation is a complex process, commonly involving kinetic competition between the formation of amorphous aggregates and fibrillar species, including a variety of intermediates, multiple conformational states, and a number of filamentous forms. A schematic diagram representing the various conformational states a protein can adopt during folding, misfolding and aggregation is presented in Figure 1.10.
Figure 1.10 A schematic representation of some of the many conformational states adopted by polypeptide chains during folding, misfolding and aggregation. The figure was reproduced from the detailed review presented by Chiti and Dobson [101].
The fibril forming process is believed to follow a series of steps: monomerisation, formation of partially folded intermediates, nucleation, and fibril growth. A variety of intermediate structures are formed due to hydrophobic interactions between the polypeptide chains. Detailed description of the proposed mechanisms and models of fibril formation is presented below.

**Nucleation**

Amyloid fibril formation has many characteristics of a “nucleated growth” mechanism. The time course of the conversion of a peptide or protein into its fibrillar form, experimentally measured by ThT fluorescence or light scattering, typically includes a lag phase that is followed by a rapid exponential growth phase [101]. The lag phase is assumed to be the time required for nuclei to form. Once a nucleus is formed, fibril growth is thought to proceed rapidly by further association of either monomers or oligomers with the nucleus. As with many other processes dependent on a nucleation step (including crystallisation), the addition of preformed fibrillar species to a sample of protein under aggregation conditions, referred to as “seeding”, causes the lag phase to be shortened and ultimately abolished. At this stage the rate of the aggregation process is no longer limited by the need for nucleation [113].

All the models proposed so far for fibril formation involve significant conformation changes during the fibrillation process. Nelson and Eisenberg suggested that there is no single model that accounts adequately for all properties of all fibrils, however there are several models that provide valuable insight into the conversion of proteins from soluble to their fibrous forms [114]. Some of the models proposed are: refolding, natively disordered and gain-of-interaction.

**Refolding**

Refolding models depict each fibril-forming protein as existing in either its native state or a distinctly different fibril state. In the conversion from native to fibril structure, the protein must unfold and then refold. Because the fibrillar state has many common properties that are independent of the fibril-forming protein, such as morphology and diffraction pattern, it has been suggested that the fibrillar state is defined by backbone hydrogen bonds [115]. In this model, the specific sequence of amino acid side-chains is unimportant, although the composition can affect the rate of fibrillisation and the stability of the fibrillar state [116]. An
example of a refolding model is that proposed for insulin by Jiménez et al. who based their model on a cryo-EM reconstruction of insulin fibrils [117].

**Natively disordered**

The natively disordered model suggests that proteins or segments of proteins, which are poorly ordered in their native states, form fibrils by becoming structured to form the cross-$\beta$ spine. This model has been proposed for several natively disordered proteins ([114] and the references therein).

**Gain-of-interaction**

The third general class of model is termed gain-of-interaction. In this model a conformational change in a limited region of the native protein exposes a previously inaccessible surface. This newly exposed surface binds to another molecule, building up a fibril. In the gain-of-interaction model, most of the structure of the native protein is retained in the fibril, only the interaction surface and its links to the core domain are changed.

The gain-of-interaction models can be further divided into four classes. The first class is called the direct stacking model, which suggests that the newly exposed surface attracts a complementary surface of an identical molecule, which will then stack on top of each other to form a fibril. This model of fibril formation is least compatible with the cross-$\beta$ diffraction patterns, thus the generality of the direct stacking model for all amyloid-like fibrils seems unlikely.

Cross-$\beta$ spine models are a second class of the gain-of-interaction model. In these models, the short segment of the protein chain that becomes exposed has a tendency to stack into a $\beta$-sheet. The fibril grows with the stacking of the short segments of many identical molecules into $\beta$-sheets. The segment may be located at the end of a folded domain or between two folded domains, and in either case the domains are proposed to retain their native structure in the fibril.

The last two classes of gain-of-interaction models are the three-dimensional domain swapping and the three-dimensional domain swapping with a cross-$\beta$ spine model. As suggested by the names, the structured domains of the protein can swap position and stack to form fibrils with or without a cross-$\beta$ spine. These models have been proposed for fibrils of cystatin C, $\beta_2$-microglobulin and others, see references within [114].
Oligomer formation and fibril growth

It has been proposed in several studies [118-120] that the association of monomers to the amyloid fibril occurs by multi-step kinetics, based on the elongation of the fibril by a dock-lock mechanism [119, 121]. In this model, an initial reversible weak “docking” interaction is required to enable subsequent “locking” interactions to occur between the formed oligomer or fibril and the deposited monomer. The “locking” event is the rate limiting step in this kinetic model of fibril formation.

It is important to understand the mechanisms of fibril formation starting from the monomeric state because it is becoming increasingly evident that the nonfibrillar intermediates may be the toxic species in amyloid related diseases. This has been shown for Alzheimer’s disease [122]. Experimental characterisation of the formation of oligomers and their structures is difficult to perform because of their diverse morphologies and rapid conformational fluctuations [123], so this characterisation has been mostly tackled by computational techniques [112, 124-126].

The mechanisms and rates of fibril formation can be affected by many factors. They may be enhanced or reduced by the local environment, such as changes in the metal ion concentrations, temperatures, pH conditions, organic solvents, or cosolvents. Other factors include mutations, transmitted prion proteins, or simply the inevitable aging process. In the following sections a review on some current studies that have investigated the influences of mutation, lipids and oligomeric stability on the propensity for fibril formation is presented.

1.4.5 Sequence and mutation effects on fibril formations

The sequence and location of the fibril forming residues is important, as their discovery will add to the understanding of the fibril forming process of the characteristic protein and to the development of appropriate therapeutics. Hydrophobic effects can be considered to be the main driving force determining aggregation and fibrillation. Recently, there has been significant interest in residue specific factors such as the number of aromatic side-chains and clustering of hydrophobic residues, exposed surface area, and effect of dipole moment on the propensities for fibril formation.

Amino acid substitutions within regions of the sequence that play a crucial role in the behaviour of the protein can reduce or increase the rates of aggregation. Residue mutations
have an effect on amyloid formation only when they are located in specific regions of the sequence [127, 128]. Mutational analysis methods have been employed to predict the effect of mutations on the aggregation rates of a wide variety of polypeptide chains [129]. More recently, the role of the aromatic rings in fibril formation was discovered to be significant, specifically in relation to their presence in fibril forming regions and their contribution to fibril stability [130]. Site-directed modification performed on short amyloidogenic fragments revealed that aromatic residues play a crucial part in the fibrillation process [131].

The increasing knowledge of such effects is beginning to lead to an understanding of the factors that cause only specific segments of the sequence to form the characteristic cross-β structure. Based on these discoveries algorithms have been developed to identify the regions of the sequence that are likely to promote aggregation within a polypeptide chain [132, 133]. The outcome of both these approaches is a plot of aggregation propensity as a function of residue number. These plots have shown to be in a very good agreement with the experimentally determined sequences which promote and stabilise the aggregation in Aβ peptide and α-Synuclein. Additional literature review on the effects of oxidation and mutation of amyloidogenic peptides is presented in Chapter 6.

1.4.6 Lipid effects on fibril formation

Recent studies have revealed a number of pathophysiological features which appear to be common to many amyloid related conditions. One such common feature is the involvement of lipids at various stages of the disease evolution. The lipid-associated peptide toxicity and aggregation enhancement has been widely established under a variety of lipid models such as micellar [134] and bilayer membranes [135], even free fatty acids and lipids [136, 137].

Lipid-protein interactions play a key role in a wide variety of cellular processes, including signal transduction, intracellular transport, enzyme catalysis and control of membrane fusion [138]. However, studies have shown that protein and peptide aggregation is enhanced in a membrane environment [139]. Amyloid formation has been reported to induce membrane permeabilisation resulting from alterations in the bilayer structure and/or uptake of lipids into the forming fibril [140]. It has also been hypothesised that the extraction of membrane lipids by the forming amyloid may be the direct cause for membrane permeabilisation and cell death [138]. The fibril forming mechanisms on lipid bilayers or membranes are similar to the ones in solution, however, recent studies have alluded to the
substantial accelerating effect lipids and membranes have on fibril formation. This is most likely due to the favourable surface effects for protein aggregation.

Recently, experimental studies have shown various behaviour of fibril formation in the presence of free lipids [141, 142]. In particular, fibril inhibition and formation was observed to be dependent on the lipid concentration. The effects of micellar and sub-micellar lipid concentrations on the mechanisms of fibril formation have been investigated for several proteins. Zhu and Fink investigated the interaction of α-Synuclein with lipid vesicles of various sizes and concentrations [143]. They found that at high peptide:lipid ratio of 1:5, α-helical structures were induced which inhibited fibril formation. At lower lipid concentrations, partially folded conformations were observed which led to fibril formation. The presence of phospholipids below and above the critical micelle concentration (CMC) has also been demonstrated to alter the kinetics of apolipoprotein C-II fibril formation and morphology [142, 144].

The presence of lipids and lipid complexes in aggregated β-sheet amyloid fibrils is well documented (additional literature is presented in Chapter 7). However, their role in the fibril forming process, specifically the involvement of the lipid-peptide interaction in the fibril formation or inhibition is still poorly understood.

### 1.4.7 Oligomer stability

As discussed in the previous sections, more recent findings have raised the possibility that precursors to amyloid fibrils, such as low-molecular-weight oligomers, and/or structured protofibrils, are the real pathogenic species. The most documented evidence is that of the Aβ protein found in the brains of patients with Alzheimer’s disease. The severity of cognitive impairment in Alzheimer’s disease correlates with the levels of low-molecular-weight species of Aβ, including small oligomers, rather than the full grown fibrils [145]. Genetic evidence also supports the theory that precursor aggregates, as opposed to mature fibrils, are the pathogenic species. The aggressive “Arctic” (E693G) mutation of the amyloid β precursor protein, associated with a heritable early-onset manifestation of Alzheimer’s disease, has been found in vitro to enhance protofibril, but not fibril formation [146].

The small size and ability to investigate these structures under controlled conditions have made them the subject of many computational studies [112, 124-126]. In particular, the ability to determine specific conformational features and interactions that influence the
CHAPTER 1. Computational studies of protein folding and aggregation

formation and stability of these soluble oligomers is of great interest. Ma and Nussinov investigated the stability of several Aβ fragments under high temperature simulation conditions. Their simulations showed that an anti-parallel β-sheet orientation is the most stable for the Aβ16-22 peptides, in agreement with a solid-state NMR based model [112]. The propensities for peptide dimerisation by Aβ10-35 was investigated by Tarus et al. using umbrella sampling and classical MD simulations [147]. Their results showed that the structure of the peptide dimer is determined by the favourable desolvation of the hydrophobic residues at the interface, and is stabilised by hydrophobic interactions. Melquiond et al. studied the aggregation paths of seven chains of the shortest amyloid-forming peptide KFFE, using an activated method and a reduced atomic representation. They found disordered KFFE monomers form three distinct topologies of similar energy: amorphous oligomers, incomplete rings with β-barrel character, and cross-β structures.

The full elucidation of the aggregation process of a protein requires the identification of all the conformational states and oligomeric structures adopted by the polypeptide chain during the fibril forming process. In particular, the determination of the thermodynamics and kinetics of all the conformational changes that link these different species are of great importance. The characterisation of each transition in molecular detail and the identification of the residues and regions of sequence that promote fibrillation are crucial for the development of therapeutic methods. Furthermore, the identification and characterisation of the oligomeric structure preceding the formation of fibrils is of particular interest because of an increasing awareness that these species are likely to play a critical role in the pathogenesis of amyloid diseases. The urgent need for elucidation of the molecular mechanisms of protein folding, misfolding and aggregation under various conditions highlighted in this chapter is directly linked with the aims of this project.
Chapter 2

2. Computational techniques for protein studies

2.1 Overview

Computational methodologies are increasingly becoming accepted as a complementary technique to experiments. The increase in computer power and improvement in theoretical algorithms has enabled computational modelling to be applied to study the physical basis of structure and function of biological macromolecules.

This chapter presents an overview over some of the more frequently used procedures for investigating protein function and dynamics under various conditions. Their application in the studies presented in this thesis is also noted. Computational techniques such as molecular dynamics and their underlying methodologies that allow simulations to be a projection of real atomic events are discussed in Sections 2.4 and 2.5, respectively.
Molecular dynamics (MD) simulations were utilised to investigate several aspects of protein folding and dynamics. First, the systematic force field effect on the structure and dynamics of proteins (Chapter 4) was studied. The effects of mutations and lipid environment on the dynamics of fibrillogenic peptides were studied and the findings are presented in Chapter 6 and 7, respectively. Furthermore, the structural stability of pre-formed oligomeric composites of the fibrillogenic peptide was examined using MD and the results are discussed in Chapter 8 of this thesis.

Complications due to insufficient conformational sampling are examined in Section 2.5, along with an introduction to some recently developed techniques which provide possible solutions to the sampling problem. The methodologies underlying Umbrella Sampling and Bias-Exchange Metadynamics (BE-META), which are methods capable of exploring wider conformational space of proteins, are introduced in Sections 2.6 and 2.9, respectively.

The novel technique BE-META was implemented to effectively sample the conformational space of insulin. Specifically, the folding pathways of insulin chain B were investigated. Using this method and complementary analysis algorithms the thermodynamic and kinetic properties of our protein model was obtained. The results are discussed in detail in Chapter 5.

The Umbrella Sampling method, together with the Weighted Histogram Analysis Method (WHAM) was applied to investigate the aggregation energetics of fibrillogenic peptides under various conditions. Specifically, the effect of mutations on the dimerisation of the peptides, as well as the effect of lipids and their concentrations on the peptide conformation was studied. The methodology applied and results from these calculations are presented in Chapter 7 and Chapter 8, respectively. Finally, in Section 2.10 the approach taken for the construction of our protein models and the detailed procedures applied for each protein simulation are described.
2.2 Molecular modelling

Molecular modelling is a collective term for theoretical methods and computational techniques that are applied to model or mimic the behaviour of molecules, ranging from small chemical systems to large biological molecules or material assemblies. While experiments are useful in determining many properties of different materials, they are still very limited in studying phenomena at atomic resolution within short time scales at various environmental conditions. An advantage of computer simulations over traditional experimental methods is the ability to characterise the time evolution of a system under controlled conditions at atomistic detail.

Molecular modelling of proteins was initiated by Corey and Pauling [148] with models originally made of hard wood (1 inch per Å) and plastic (0.5 inch per Å). The wood models were connected with steel rods and clamps, the plastic models with snap fasteners. This model was based on structural data obtained from X-ray diffraction (XRD) and used correct atom proportions based on their van der Waals radii. It was Koltun in 1960 who improved the original Corey-Pauling model, resulting in the Corey-Pauling-Koltun (CPK) model which has proven to be very useful in visualising and making accurate measurements of protein structure.

Today computer simulation techniques can be classified into several categories based on the studied system size and time-scale capabilities, as shown in Figure 2.1. The most accurate methods are based on quantum mechanics (QM), and comprise the \textit{ab-initio} and Density Functional Theory (DFT) methods. Protein models are too large to be treated by these techniques as the properties of a system are calculated from first principles, i.e. considering the electronic structure, which is computationally prohibitive for proteins. A less expensive approach is offered by the \textit{semi-empirical methods} which allow for larger systems to be simulated at the electronic level, treating only the valence electrons explicitly and fitting several parameters to experimental data. However, the QM methods are mostly used for investigating processes which involve bond breaking and formation, electronic rearrangement and force field parameterisation.
Larger system sizes and longer simulation time scales can be investigated using the atomistic force field methods or molecular mechanics. The main concept behind this method is the *Born-Oppenheimer approximation*, which allows for the expression of the electronic movement to be eliminated from the Hamiltonian of the system, leaving only the nuclear variables. Molecular mechanics is used when physical interactions are considered (as opposed to chemical) and based upon a simpler model of interactions within a system with contributions from processes such as stretching of bonds, the opening and closing of angles (angle bending) and rotations about single bonds (torsions). This method can provide accurate calculation of physical phenomena such as protein dynamics, in a fraction of the computer time compared to *ab-initio* calculations. All the calculations performed in this thesis have utilised molecular mechanics-based methods. Detailed discussion of the principles and force fields behind this powerful technique is presented in Section 2.3 of this chapter.

Additionally, coarse-grained methods are receiving significant focus at the moment due to their ability to simulate large complexes at timescales currently inaccessible to atomistic methods. These methods can also provide qualitatively accurate representation of the macroscopic states and some properties of a system over long timescales. In coarse-grained models a small group of atoms is treated as a single interaction unit or bead, where
the dynamics of the system is governed by a very simple force field. There is an intrinsic
difficulty in the parameterisation of coarse-grained force fields, related to the fact that
complex and diverse interactions must be described by a small number of parameters.
However, continued development and improvements of different coarse-grained models
means this method has great potential for modelling large complex systems comparable with
experiments. Examples of possible applications of coarse-grained methods as a continuation
of this project are discussed in the Future work chapter of this thesis.

2.3 Molecular mechanics: Empirical force field models

Empirical force field methods are important tools for the application of theoretical approaches
to investigate structure-activity relationships in biological systems at atomistic resolution
[149]. The key advantage of these methods and reason for their continuous use is their
accuracy, defined by their ability to reproduce properties accessible by experiments.

2.3.1 Potential energy function

Fundamental to molecular mechanics simulations are the forces that govern the atomic
motions, derived from pairwise atom-atom interactions usually given in the form of an
empirical potential energy function. The combination of the potential energy function with the
geometric and energetic parameters used, yield what we know as a force field (FF). The most
commonly used force fields for biomolecular simulations are the all-atom CHARMM [150],
AMBER [151], OPLS-AA [152] force fields, where all the atoms in the proteins are
represented explicitly. There are also the united-atom GROMOS [153] force fields, where the
hydrogens bonded to an aliphatic carbon are treated as a single interaction site, thus the name
“united-atom”. A brief introduction to these FFs’ functional forms and parameterisation is
discussed below. There are other force fields available for protein simulations which include
CVFF [154], ECEPP [155, 156], ENCAD [19, 157], MM4 [158], MMFF [159, 160], and UFF
[161], however we will limit our description to the preceding most commonly used and well
tested four force fields.

The underlying functional form of each of these force fields contains energy terms
describing the bonded ($E_{\text{bonded}}$) and nonbonded ($E_{\text{nonbonded}}$) interactions between the atoms of a
system. The terms representing bonded interactions account for the stretching of bonds, the bending of valence angles, and the rotation of dihedrals. The terms representing the nonbonded interactions define the electrostatic and van der Waals interactions.

\[ E_{\text{bonded}} = \sum_{\text{bonds}} K_b (b - b_0)^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_0)^2 + \sum_{\text{dihedrals}} K_\chi (1 + \cos(n\chi - \sigma)) \]  \hspace{1cm} (2.1)

and

\[ E_{\text{nonbonded}} = \sum_{\text{nonbonded pairs}} \left( \varepsilon_{ij} \left( \frac{R_{\text{min},ij}}{r_{ij}} \right)^{12} - 2 \left( \frac{R_{\text{min},ij}}{r_{ij}} \right)^{6} \right) + \frac{q_i q_j}{r_{ij}} \]  \hspace{1cm} (2.2)

The total potential energy function is in the form of:

\[ E_{\text{total}} = E_{\text{bonded}} + E_{\text{nonbonded}} + E_{\text{other}} \]  \hspace{1cm} (2.3)

where \( E_{\text{bonded}} \) is the contribution to the total energy from the bonded interactions, \( E_{\text{nonbonded}} \) is the contribution from the nonbonded interactions, and \( E_{\text{other}} \) includes any force field specific terms, which will be discussed subsequently.

The first term in Equation 2.1 describes the stretching of bonds in a quadratic form, equivalent to that of Hooke’s law for the potential energy of a spring. The equation is a sum over all bonded pairs of atoms, where \( b \) is the interaction distance or bond length, \( K_b \) and \( b_0 \) are the parameters describing the stiffness and the equilibrium length of the bond, respectively. The second term involves triplets of consecutively bonded atoms and describes the bending of the angle (\( \theta \)) they are forming. \( K_\theta \) and \( \theta_0 \) are the parameters describing the stiffness and equilibrium geometry of the angle, respectively. Similar to the bond stretching term this equation also has a quadratic form. The third term from the bonded interaction energies equation describes the energetics associated with rotation of the dihedral angle defined by quadruplets of consecutively bonded atoms. As dihedral rotations are periodic in nature, a cosine function is used, where \( \chi \) is the value of the dihedral, \( K_\chi \) is the energetic parameter that determines the barrier height, \( n \) is the periodicity and \( \sigma \) is the phase.

Equation 2.2 describes the intermolecular or nonbonded interactions. In the four widely used force fields mentioned above, nonbonded interactions between atoms are defined
as occurring either between atoms in separate molecules or between atoms separated by three or more bonds in the same molecule. The first part of Equation 2.2 is the van der Waals term, in the form of a Lennard-Jones equation, and models the attractive dispersion and repulsive interactions. The prefactor $\varepsilon_{ij}$ is a parameter based on the types of the two interacting atoms $i$ and $j$. $R_{\text{min},ij}$ is a parameter that also depends on the types of the two interacting atoms and defines the distance at which the Lennard-Jones energy is minimum. The second part of Equation 2.2 models the electrostatic interactions between nonbonded pairs of atoms and is based on Coulomb’s law. As with the Lennard-Jones equations, $r_{ij}$ is the interatomic distance, while $q_i$ and $q_j$ are the parameters that describe the effective charges on atoms $i$ and $j$. These charges are partial atomic charges with noninteger values that are selected to represent the overall charge distribution of a molecule. Even the hydrogen atoms on aliphatic carbon atoms can have charges of approximately 0.05 to 0.1 electrons in biomolecular force fields. Naturally, the sum of the partial charges in a molecule must equal the molecule’s net formal charge.

Having concluded the discussion of a generic force field, there are a few variations between the individual force fields worthwhile mentioning. There are two primary differences in the bonded portion of the potential energy functions. The first is the variable of “improper” dihedrals, which can be applied to maintain chirality or planarity at an atom centre bonded to three other atoms, e.g. for cyclic molecules like benzene. In the case of AMBER and OPLS-AA, improper dihedral angles contribute to the energy via the dihedral term in Equation 2.1, and are applied to planar groups with a periodicity $n = 2$. The CHARMM and GROMOS force field add a separate term for improper dihedral energy that has a quadratic dependence on the value of the improper dihedral, similar to the terms for bonds and angles. This is particularly important for the GROMOS force field because as a united-atom force field, it does not include particle positions for hydrogen atoms bonded to aliphatic carbons, so the improper dihedral terms serve to preserve chirality at these carbon centres. The second difference between the functional forms of the four abovementioned force fields is that the CHARMM force field adds a Urey-Bradley angle term. This formulation treats the two terminal atoms in an angle with a quadratic term that depends on the atom-atom distance. The improper dihedral and Urey-Bradley angle terms provide additional degrees of freedom for the accurate reproduction of vibrational spectra during parameterisation. Differences also exist between the nonbonded terms of the four force fields. One such difference is concerned with the various scaling constants each force field applies to the Lennard-Jones and Coulomb
interactions between the atom pairs. A detailed description of the differences between the energy terms of various force fields is provided in the article by Guvench and MacKerrell [162].

### 2.3.2 Force field parameterisation

The energy functions discussed above are not of any value if they are not accompanied with a set of parameters that describe the energetic and geometric properties of the interacting particles. The optimisation of force field parameters involves adjusting the parameter values so the force field is able to reproduce experimental data. This may include the use of experimental spectroscopic, thermodynamic, and crystallographic data as well as data computed using quantum mechanics methods. Typical examples of experimental data used for parameterisation of a force field or the consequent refinement are: vibrational spectra, densities, solvation free energies, electron, or X-ray diffraction structures, and relative conformational energies and barrier heights. The AMBER, CHARMM, OPLS-AA and GROMOS force fields are each based on a different type of experimental data, although there is some overlap. The parameters for these force fields were extensively optimised with particular emphasis on the treatment of proteins. With OPLS and CHARMM22 the partial atomic charges were based on Hartree-Fock/6-31G* supramolecular data [150, 163] while the standard AMBER release (AMBER99) is based on restrained electrostatic potential (RESP) charges fit to the same level of theory. There are several variations available of the AMBER force field. Duan et al. extensively modified the previously available AMBER94 [151] and AMBER99 [164], now called AMBER03, in which the derivation of the partial atomic charges was fundamentally different [165]. A low-dielectric continuum model corresponding to an organic solvent environment was directly included in the quantum mechanical (QM) calculation of the dihedral parameters and electrostatic potential, from which the charges are obtained. In contrast to the parameterisation of other biomolecular force fields, the parameterisation of the united-atom GROMOS 53A6 force field is based primarily on reducing the free enthalpies of hydration and apolar solvation for a range of compounds.

### 2.3.3 Treatment of solvation

An essential aspect of the use of a force field for biomolecular simulations is the accurate treatment of the condensed aqueous environment. When selecting a water model to use for a
particular study, the most important consideration is its compatibility with the biomolecular force field used. The reason for this is that most force fields have been developed in conjunction with a specific water model. For example, AMBER, CHARMM and OPLS have been developed with the TIP3P water model, OPLS also with TIP4P while GROMOS with SPC and F3C models. The most commonly used water models are TIP3P and SPC models. The limitation of the TIP3P water model is that it underestimates the height of the tetrahedral peak in the O–O radial distribution function and that the diffusion constant is larger than the experimental values [166]. However, this model does treat energetics satisfactorily and because the majority of biomolecule-water interactions involve first or second shell hydration, the lack of long-range structure is often not relevant to the analysis. The SPC model is similar to TIP3P, although by applying tetrahedral geometry, H–O–H angle of 109.47°; it has slightly increased long-range structure, as evidenced by a more defined tetrahedral peak in the O–O radial distribution function. An extensive study of hydration thermodynamic properties on several side-chains, using AMBER99, GROMOS 53A6 and OPLS, with five water models, SPC, SPC/E, TIP3P, TIP4P, and TIP4P-Ew (see Hess and van der Vegt [167] and the references within) identified only small differences in hydration free energies between the water models, however TIP3P performed slightly better with AMBER99 and OPLS than the other water models. The GROMOS 53A6 force field showed an overall better performance than the other force fields for the free energy, which is expected, as it was parameterised against this property.

### 2.3.4 Force field validation studies

There are many articles which have performed force field validation calculations and consecutively performed their own parameter corrections. Gnanakaran and Garcia identified that all-atom protein force fields that predict reasonable conformational dynamics for larger peptides or proteins, fail to reproduce the measured conformational distribution for di- and tripeptide systems. Specifically, CHARMM22, AMBER94 and AMBER96 are reported as having such defects [168].

Even though many of the force fields used currently give rise to three main conformations for short peptides, α-helix, β-strand and coil, the relative population of these structural elements varies significantly, as discovered by Mu et al. [169] and Hu et al. [170]. The CHARMM22 force field has a strong preference for α-helical conformations for di- and
CHAPTER 2. Computational techniques for protein studies

tri-peptides [169-171], whereas 2D-IR and NMR measurements consecutively show that these peptides adopt primarily polyproline II (PP$_{II}$) conformations [169]. The AMBER94 force field has been shown to favour $\alpha$-helical and AMBER96 favours extended structures for both small and large proteins. However, AMBER96 inadequately models the thermodynamics of helical peptides [172] as well as folding of turn structures [173]. OPLS-AA does not separate the PP$_{II}$ and $\beta$ structure basins in the Ramachandran map, while GROMOS96 adequately samples $\beta$ conformations, but is less effective in sampling of the $\alpha$-helical elements [169, 170]. It is evident from all the studies summarised that force fields require continuous improvement, parameterisation and validation, so that true representations of protein behaviour can be obtained.

In order to identify the structural anomalies a force field may exert on a protein, a systematic comparison was performed of the most widely used and generally available force fields for studying biomolecules, CHARMM27 [150], OPLS-AA [163], AMBER03 [165], GROMOS 43A1 [174] and GROMOS 53A6 [153]. The different force field effects on our model system chain B of insulin was investigated and the findings of this study are presented in Chapter 4 of the thesis. Additional force field validation studies are also presented.

2.4 Molecular dynamics: Theory and application to protein simulations

The molecular dynamics simulation technique is one of the most important and widely used methods for studying the structural time evolution of many particle systems at atomic detail. The first molecular dynamics application was performed by Alder and Wainright [44, 175] on a system of hard spheres, as a model for atomic-scale systems. This was the first application of molecular dynamics and since then it has been applied in a variety of systems, including polymers, nano-materials and different biological composites of various size. A comprehensive review on some of the application of molecular dynamics can be found in reference [176].
2.4.1 Equations of motion

The molecular dynamics simulation method is based on Newton’s law of motion, $F = ma$. By solving the differential equations embodied in Newton’s second law, we are able to obtain a trajectory that describes the positions, velocities, and accelerations of the particles as they progress over time.

$$ F_i = -\frac{dE}{dr} \quad (2.4) $$

$$ \frac{F_i}{m_i} = \frac{d^2r_i}{dt^2} \quad (2.5) $$

The force $F_i$ acting on the particles in the system is derived from the potential energy $E(r^N)$ (Equation 2.3) which is defined by the selected force field, where $r^N = (r_1, r_2, r_3, ..., r_N)$ represents the complete set of $3N$ atomic coordinates. Even though the equation is simplistic in its form, there is no analytical solution to the equation of motion for systems of more than two particles, so it must be solved numerically. Numerous algorithms have been developed for solving these equations, and most are derived from the widely used Verlet algorithm [177], with variations including the leap-frog algorithm [178], velocity Verlet [179] and Beeman’s algorithm [180]. All algorithms assume that the positions and dynamic properties can be approximated as Taylor series expansions. The Verlet algorithm requires the knowledge of current positions, $r(t)$; acceleration $a(t)$; and the position from the previous step, $r(t-\delta t)$. The position of the next step can then be found using the formula:

$$ r(t+\delta t) = 2r(t) - r(t-\delta t) + \delta t^2a(t) \quad (2.6) $$

The velocities do not explicitly appear in the Verlet integration algorithm, however they can be calculated by a simple approach to define the difference in positions at time $t+\delta t$ and $t-\delta t$ by $2\delta t$ resulting in:

$$ v(t) = \frac{r(t+\delta t) - r(t-\delta t)}{2\delta t} \quad (2.7) $$

Some of the advantages of this method are that it is straightforward and has low storage requirements, however it is an algorithm of moderate precision. The lack of an explicit
velocity term in the equations makes it difficult to obtain the velocities, and indeed the velocities are not available until the positions have been calculated at the next step.

A variation on the Verlet algorithm is the leap-frog integration technique. This method uses the following relationships:

\[
\mathbf{v}\left(t + \frac{1}{2} \delta t\right) = \mathbf{v}\left(t - \frac{1}{2} \delta t\right) + \delta t \mathbf{a}(t) \tag{2.8}
\]

\[
\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta \mathbf{v}\left(t + \frac{1}{2} \delta t\right) \tag{2.9}
\]

In this algorithm, the velocities \( \mathbf{v}(t + \frac{1}{2} t) \) are first calculated from the velocities at time \( t - \frac{1}{2} \delta t \) and the accelerations at time \( t \) (Equation 2.8). The positions \( \mathbf{r}(t + \delta t) \) are then deduced from the velocities just calculated together with the positions at time \( \mathbf{r}(t) \) using Equation 2.9. The velocities at time \( t \) can be calculated by the relationship:

\[
\mathbf{v}(t) = \frac{1}{2} \left[ \mathbf{v}\left(t + \frac{1}{2} \delta t\right) + \mathbf{v}\left(t - \frac{1}{2} \delta t\right) \right] \tag{2.10}
\]

The velocities thus ‘leap-frog’ over the positions to give their values at \( t + \frac{1}{2} \delta t \). The positions then leap over the velocities to give their new values at \( t + \delta t \), ready for the velocities at \( t + \frac{3}{2} \delta t \), and so on (hence the name leap-frog). The advantage of this algorithm is that the velocities are explicitly calculated, however, the disadvantage is that they are not calculated at the same time as the positions.

### 2.4.2 Periodic boundary conditions

One of the limitations of computational methods is the inability to effectively simulate large systems over long periods of time, although sometimes this is necessary depending on the problem to be investigated. The correct treatment of boundaries and border effects is crucial to simulation methods because it enables “macroscopic” properties to be calculated from
simulations using a relatively small number of particles. The traditional way to minimise the
edge effects in a finite system is to apply periodic boundary conditions in such a way that the
particles experience forces as if they were in bulk fluid. The atoms of the system to be
simulated are placed into a space-filling box, which is surrounded by translated copies of
itself. A two-dimensional illustration of applied periodic boundary conditions is shown in
Figure 2.2. In the two-dimensional example each box is surrounded by 8 neighbours; while in
three dimensions each box would have 26 nearest neighbours. The coordinates of the particles
in the image boxes can be computed by adding or subtracting integral multiples of the box
sides. If a particle leaves the box during the simulations, it will be replaced by an image
particle that enters from the opposite side. The number of particles within the central box thus
remains constant.

![Figure 2.2 A two-dimensional illustration of periodic boundary conditions.](image)

Depending on the system to be examined, there are several space-filling type boxes
that may be more computationally efficient for running molecular dynamics calculations.
There is the cubic unit cell, rhombic dodecahedron, truncated octahedrons, and the most
general space-filling shape, the triclinic unit-cell [181]. A typical triclinic simulation box
containing a model protein in explicit solvent is presented in Figure 2.3.
2.4.3 Neighbour search and nonbonded interactions

The most time consuming part of a molecular dynamics simulation is the calculation of the nonbonded energies and forces, also known as the long-range electrostatic and Lennard-Jones effects. The most popular approach to deal with the nonbonded interactions is to apply the minimum image convention [182] where the nearest image of each particle is considered for calculation of the short-range nonbonded interaction terms. Another approach is to implement the nonbonded cutoff or potential truncation method. When a cutoff is employed, the interactions between all pairs of atoms that are further apart than the cutoff value are set to zero. This is because the greatest contribution to the potential and forces comes from the neighbouring particles, therefore only the closest image is taken into account. The potential truncation method introduces a small perturbation to the potential and force calculations, rendering it not accurate enough for calculations of long-range electrostatic interactions, and can become computationally expensive as the system size increases. To lessen the computational burden, various truncation schemes have been developed whereby nonbonded interactions beyond a cutoff distance are ignored, and a smoothing function is typically applied to ensure continuity in the forces [183]. Such procedures imply severe approximations in the case of electrostatic interactions with energy decaying as $1/r_{ij}$. As a result, these interactions continue to make a substantial contribution to the total system energy beyond the
typical cutoff length of approximately 10 Å. When periodic boundary conditions are being used the cutoff should not be so large that a particle sees its own image or the same molecule twice. This has the effect of limiting the cutoff to no more than half the length of the cell when simulating atomic fluids.

The introduction of Ewald sums [184] into biomolecular simulations under periodic boundary conditions has solved some of the problems associated with cutoff methods. The system is treated as being infinitely periodic, and interaction energies and forces beyond the cutoff length are calculated using the Ewald algorithm. This method works in reciprocal space, instead of being calculated directly using Coulomb’s law. Although the original Ewald formalism is computationally expensive, recent developments based on grid-based treatments of reciprocal space, including Particle-Mesh Ewald (PME) methods [185], allow for the rigorous treatment of long-range electrostatics in a computationally efficient manner [186, 187].

### 2.4.4 Thermodynamic ensembles

The simplest MD simulation of a system under periodic boundary conditions is evolved under constant energy and constant volume conditions (NVE). Although thermodynamic results can be transformed between ensembles, this is strictly only possible in the limit of infinite system size, also known as “the thermodynamic limit”. It may therefore be desired to perform the simulation in a different ensemble, such as the canonical ensemble (NVT) and isothermal-isobaric ensemble (NPT). Additionally, many experimental measurements are made in environments of constant temperature and pressure, and so simulations in the isothermal-isobaric (NPT) ensemble are more directly relevant to experimental data.

### 2.4.5 Temperature coupling

For several reasons such as solute drift during equilibration, drift as a result of force truncation and integration errors, heating due to external or frictional forces, it is necessary to control the temperature of a system. The weak coupling scheme of Berendsen [188] and the extended-ensemble Nosé-Hoover scheme [189, 190] are the two most widely used temperature coupling methods.
The Berendsen temperature coupling algorithm mimics weak coupling with first-order kinetics to an external heat bath with a given temperature $T_0$. The effect of this algorithm is that a deviation of the system temperature from $T_0$ is slowly corrected according to

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau}$$

(2.11)

which means that a temperature deviation decays exponentially with a time constant $\tau$. This method of coupling has the advantage that the strength of the coupling can be varied and adapted depending on the simulation requirements. For example, for equilibration purposes the coupling time can be set quite short (e.g. 0.01 ps), but for a reliable statistics collection stage it can be set much longer (e.g. 0.5 ps), with minimum influence on the system’s dynamics. The heat flow into or out of the system is effected by scaling the velocities of each particle every step with a time-dependent factor $\lambda$, given by

$$\lambda = \left[1 + \frac{\Delta t}{\tau_T} \left(\frac{\frac{T_0}{T(t - \frac{\Delta t}{2})} - 1}{\frac{T_0}{T(t - \frac{\Delta t}{2})}}\right)\right]^\frac{1}{2}$$

(2.12)

The $\tau_T$ is close to, but not exactly equal to the time constant $\tau$ of the temperature coupling. This is because the kinetic energy change caused by scaling the velocities is partially redistributed between the kinetic and potential energy and hence the change in temperature is less than the scaling energy. In practice, the scaling factor $\lambda$ will always be closer to 1. The Berendsen weak coupling algorithm is extremely efficient for relaxing a system at a target temperature.

A method which gives correct description for canonical ensemble simulations, is the extended-ensemble approach first proposed by Nosé [189] and later modified by Hoover [190], now known as the Nosé-Hoover temperature coupling algorithm. In this method, the system Hamiltonian is extended by introducing a thermal reservoir and a friction term in the equations of motion. The friction force is proportional to the product of each particle’s velocity and friction parameter $\zeta$. This friction parameter, or ‘heat-bath’ variable, is an independent dynamics quantity with its own equation of motion, where the time derivative is
calculated from the difference between the current kinetic energy and the reference temperature. In Hoover’s formulation, the particles’ equations of motion from Equation 2.5 are replaced by

$$\frac{d^2 r_i}{dt^2} = \frac{F_i}{m_i} - \xi \frac{dr_i}{dt}, \quad (2.13)$$

where the equation of motion for the heat bath parameter $\xi$ is given by

$$\frac{d\xi}{dt} = \frac{1}{Q}(T - T_0). \quad (2.14)$$

In the equation above, $T_0$ denotes the reference temperature, while $T$ is the current instantaneous temperature of the system, and $Q$ determines the strength of the coupling.

Overall, it is important to keep in mind the difference between the weak coupling scheme of Berendsen and the Nosé-Hoover algorithm. When applying the weak coupling scheme a strongly damped exponential relaxation is obtained, while the Nosé-Hoover approach produces an oscillatory relaxation. This means that the actual time it will take to relax the system with Nosé-Hoover coupling could be several times longer than the period of oscillations selected, which is important to note in the set-up stage of molecular dynamics simulations.

### 2.4.6 Pressure coupling

With similar reasoning to that behind the need for temperature coupling, a system can also be coupled to a pressure bath to give the correct representation of a molecular system. The two most widely used methods are the Berendsen algorithm [188] and the extended-ensemble Parrinello-Rahman approach [191]. A key advantage of these methods is that they can be combined with any of the temperature coupling methods described previously.

The Berendsen pressure coupling algorithm rescales the coordinates and box vector every step with a matrix $\mu$, which has the effect of a first-order kinetic relaxation of the pressure towards a given reference pressure $P_0$:
The scaling matrix $\mu$ is given by

$$
\mu_{ij} = \delta_{ij} - \frac{\Delta t}{3 \tau_p} \beta_{ij} \left[ P_{0ij} - P_{ij}(t) \right]
$$

(2.16)

The $\beta$ value in the equation represents the isothermal compressibility of the system, where for water at 1 atm and 300 K, $\beta = 4.6 \times 10^{-5}$ bar$^{-1}$. Most other liquids have similar compressibility values. In the Berendsen algorithm the velocities of the system are neither scaled nor rotated.

For correct calculation of thermodynamic properties the fluctuations in pressure or volume are important, so it is advisable these are correctly represented in the ensemble simulated. This can be achieved using the Parrinello-Rahman approach, which is similar in form to the Nosé-Hoover temperature coupling method. With the Parrinello-Rahman barostat, the box vectors are represented by the matrix $b$.

$$
\frac{d b^2}{dt^2} = W^{-1} b^{-1} \left( P - P_{\text{ref}} \right)
$$

(2.17)

The volume of the box is denoted $V$, and $W$ is a matrix parameter that determines the strength of the coupling. The matrices $P$ and $P_{\text{ref}}$ are the current and reference pressures, respectively. The equations of motion are also changed, just as for the Nosé-Hoover temperature coupling.

$$
\frac{d^2 \mathbf{r}_i}{dt^2} = \frac{F_i}{m_i} - \mathbf{M} \frac{d \mathbf{r}_i}{dt}
$$

(2.18)

$$
\mathbf{M} = b^{-1} \left[ b \frac{db'}{dt} + \frac{db}{dt} b' \right] b'^{-1}
$$

(2.19)

Overall, by implementing the Parrinello-Rahman barostat in combination with the Nosé-Hoover thermostat an accurate isothermal-isobaric ensemble can be generated. Similar to the Nosé-Hoover thermostat, Parrinello-Rahman coupling gives oscillations of the unit-cell
vectors. Berendsen coupling on the other hand, has the advantage that the fast kinetics of the system will be unaffected and the disadvantage that it is unknown if the desired ensemble will be generated.

### 2.4.7 Bond constraint algorithms

As previously discussed, one of the most demanding aspects of simulation is the computation of nonbonded interactions because for some large systems, millions of pairs have to be evaluated for each time step. One way to enhance computational efficiency is by extending the time step used for each calculation. However, this will introduce systematic errors, as the shortest timescale in biological simulations are the hydrogen bond vibrations at 1 fs. Fortunately, in most simulations the bond vibrations are not of interest per se, and can be removed entirely by introducing bond constraint algorithms such as SHAKE [192] or LINCS [193]. Constraints make it possible to extend time steps to 2 fs.

The most widely used algorithm for large molecules is SHAKE, in which the bonds and angles are reset to prescribed values by moving the bonded particles parallel to the old bond directions. SHAKE is an iterative method, where all the bonds are reset sequentially to the correct length. Because the bonds are coupled, this procedure has to be repeated until the desired accuracy is reached. SHAKE is simple and numerically stable since it resets all constraints within a prescribed tolerance, however this method has the drawback that no solutions may be found when displacements are large. This is due to the coupled bonds being handled one by one, thus correcting one bond may tilt a coupled bond so far that the method does not converge.

A fast, reliable Linear Constraint Solver (LINCS) algorithm for molecular simulations was developed by Hess et al. where the constraints themselves are explicitly reset instead of the derivatives of the constraints [193]. The derivation of the algorithm is presented in terms of matrices; however no matrix multiplications are needed, making the method applicable to very large molecules. Although the LINCS algorithm shows the same accuracy as SHAKE, it is three to four times faster [193].
2.5 Conformational sampling problem

Molecular dynamics is a very useful and accurate method for simulating the time evolution of a system and is able to explore the free energy landscape in the neighbourhood of local minima. However, the complexity and ruggedness of the free energy surface, comprised of numerous minima separated by large energy barriers, induces difficulties in using this approach for studying complex processes such as protein folding and in some cases protein aggregation. The system can easily be trapped in one of the local minima and fail to properly sample the entire conformational space. In order to cope with this problem it is necessary to apply methodologies that are capable of accelerating rare events, specifically, configurational changes that involve the crossing of large free energy barriers.

In the last couple of decades, much time has been dedicated to the development of such methodologies, some of which have been demonstrated to be more successful than others. There are three general types of search and sampling enhancement techniques. One type of method perform deformation or smoothing of the potential energy surface, such as local potential energy elevation [194] or conformational flooding [195]. The second type consists of methods that scale the system or simulation parameters, such as temperature annealing [196] or mass scaling [197]. And the third type are multi-copy searching and sampling algorithms, such as replica-exchange and multicanonical algorithms [198]. It must be noted that there are many other techniques that fall under these categories, however only a few are mentioned here; the interested reader is referred to the articles by van Gunsteren et al. [29, 199] and references therein.

Two of the most widely used methodologies capable of exploring wider conformational space are the Umbrella Sampling [200] and Replica Exchange Molecular Dynamics [201] methods. Recently, an extension of the Metadynamics [202] algorithm was developed, called Bias-Exchange Metadynamics [63] which has shown great potential and was successfully implemented for studying the folding of chain B of Insulin [66]. A brief description of the algorithms behind these methods is presented in the following sections.
2.6 Umbrella sampling

The changes of the free energy as a function of an inter- or intra-molecular coordinate, such as the distance between two atoms, or the torsional angle around a bond within a molecule are often of interest when investigating molecular association/dissociation, conformational stretching or protein folding. The free energy surface along a chosen coordinate is known as the potential of mean force (PMF). Various methods have been proposed for calculating potentials of mean force. The simplest formulation is:

\[ A(x) = -k_B T \ln[P(x)] \]  

(2.20)

where \( P(x) \) is the probability density over the coordinate \( x \). Although this quantity may in principle be obtained from long timescale MD simulation, in practice \( P(x) \) is very slow to converge, as a large region of the PMF corresponds to a region of low occupancy. This leads to inaccurate values for the PMF, so one way around this problem is to apply a technique called umbrella sampling.

The umbrella sampling method of Torrie and Valleau [200] overcomes the problem of insufficient sampling of certain regions of the reaction coordinate \( x \) by introducing an additional biasing potential \( W(x^N) \). Using a simple example of two interacting particles, the biasing potential would force the two bodies to sample prescribed separation distances.

\[ V'(x^N) = V(x^N) + W(x^N) \]  

(2.21)

Typically, a number of simulations with different biasing potentials are carried out, each one confining the position of the particle to a particular region or “window” \( i \) on \( x \) (hence the name, umbrella sampling). A common choice for the window potentials is a harmonic one. The resultant probability density over the reaction coordinate is the biased density \( P(x') \). A common algorithm to obtain the unbiased density, \( P(x) \), and thereby retrieve the unbiased PMF, is the Weighted Histogram Analysis Method (WHAM) of Kumar et al. [203]. Illustration of several umbrella sampling windows represented as harmonic biasing potentials along a chosen reaction coordinate is shown in Figure 2.4.
CHAPTER 2. Computational techniques for protein studies

Figure 2.4 The real distribution of a system along a reaction coordinate “Distance” can be calculated using several umbrella sampling windows (red); after correcting for the bias potentials the free energy along the selected reaction coordinate is identified (purple).

2.7 Replica exchange molecular dynamics

Accurate and sufficient sampling at different temperatures is usually difficult to obtain, specifically in protein folding simulations, because of the system’s tendency to get trapped in one of the large number of local minimum energy states. The Replica Exchange Molecular Dynamics (REMD) as described by Sugita and Okamoto [201] is a powerful technique able to enhance conformational sampling, compared to running parallel simulations at constant temperatures. In this method, several noninteracting replicas of the same system are simulated in parallel, where each replica is evolved at a different temperature. At selected times a Monte Carlo temperature exchange is performed between the replicas, and an exchange is accepted or rejected based on a Metropolis acceptance criterion [182]. The Metropolis acceptance criteria states that if the difference between the energy of the resulting conformation and the energy of the current conformation is negative, then the move to another temperature is accepted. However, if the difference in energy between the two states is positive, the move will be rejected. A schematic of the temperature changes performed in an REMD simulation is shown in Figure 2.5.

High free energy barriers are crossed by the high temperature replicas while the exchanges allow the low temperature replicas to sample a larger conformational space. Only the replicas in neighbouring temperatures are probed for exchanges because the acceptance
ratio of the exchange decreases exponentially with the difference between two temperatures. If care is taken in reaching equilibrium, REMD is a powerful tool for sampling the conformational space. For this reason, REMD has wide applicability, and has been used with implicit solvent MD simulations [204] and simplified coarse-grained models [205] to study protein folding. Berne and co-workers [26] applied explicit solvent REMD simulations to obtain the free-energy landscape of a β-hairpin peptide, and Garcia and Onuchic [206] used the method to determine the free energy landscape of protein A.

While enhancing the exploration of conformational space, useful kinetic information cannot easily be extracted using the REMD method. Conformation exchanges result in discontinuous trajectories, precluding the calculation of equilibrium time correlation functions for times longer than the actual temperature exchange time. The main disadvantage of this method is that a significant number of replicas are required when a system is explicitly solvated in order to effectively sample the energy surface, thus applying this method can be computationally expensive.

![Figure 2.5](image.png)

**Figure 2.5** In replica exchange molecular dynamics simulation several replica of the same system are simulated in parallel at different temperatures. Exchanges between neighbouring replicas are performed at selected times.

### 2.8 Metadynamics

A powerful method called metadynamics [202] is a relatively new technique in which the free energy of the system is obtained from non-equilibrium simulations, and it has been successfully implemented for studying antibiotic translocation [207] and docking [208]. The metadynamics algorithm is similar to the local elevation method [194] where the system is
encouraged to visit new conformational space rather than be confined to regions of local minima. The technique is based on a dynamics performed in the space defined by several collective variables \( s(x) \), which are assumed to provide a coarse-grained description of the system, and are explicit functions of the coordinates \( x \). The dynamics of the system is biased by a history dependent potential constructed as a sum of Gaussians centred on the trajectory of the collective variables (CV):

\[
V_G(s(x), t) = \int_0^t \frac{d\tau}{\tau_G} \exp \left( -\frac{(s(x) - s(x(\tau)))^2}{2\delta^2} \right)
\]

where \( \omega \) is the height, \( \delta \sigma \) the width of the gaussians and \( \tau_G \) is the rate of their deposition. After a transient period, the Gaussian potential compensates the free energy, allowing the system to efficiently explore the space defined by the CVs. This method allows an accurate free energy reconstruction in several variables, however its performance deteriorates with increased dimensionality [202], meaning that when using greater numbers of CVs the effectiveness in conformational sampling of the method decreases, therefore limiting its usefulness for studying protein folding.

### 2.9 Bias-exchange metadynamics

A novel methodology has been recently introduced, called bias-exchange metadynamics (BE-META) [63], which does not require \textit{a priori} knowledge of the native structure. The method allows a free energy reconstruction in a virtually unlimited number of variables, and as such can be considered for investigating complex processes like protein folding [64, 65], protein-protein interactions and enzyme reactions [67]. This novel theoretical tool incorporates two previously reported powerful techniques, replica exchange [201] and metadynamics [202]. In this approach, non-interacting replicas (NR) of the system are simulated at the same temperature, each biased by a different collective variable \( s^\alpha(x) \), \( \alpha = 1, \ldots, \text{NR} \), i.e. each simulation is an individual metadynamics calculation. To achieve the greatest degree of efficiency for metadynamics, the dimensionality of each of the vectors \( s^\alpha \) is assumed to be small. Every individual replica independently accumulates a history-dependent potential.
CHAPTER 2. Computational techniques for protein studies

\( V_G^a(x, t) = V_G(s^a(x) t) \) that after a sufficiently long time, would provide an estimate of the free energy projected on \( s^a \).

The replicas are allowed to exchange the system configurations, like in the replica exchange method, where two replicas \( a \) and \( b \) are selected at random. The exchange move consists of swapping the atomic coordinates \( x^a \) and \( x^b \) of the two replicas. As they are evolved under the action of two different history-dependent potentials, the move is either accepted or rejected based on the Metropolis criterion [182].

\[
\begin{align*}
    p_{ab} &= \min\left(1, \exp\left\{ \beta \left[ V_G^a(x^a, t) + V_G^b(x^b, t) - V_G^a(x^b, t) - V_G^b(x^a, t) \right] \right\} \right)
\end{align*}
\] (2.23)

The normal potential energy of the system cancels out exactly for this kind of move. If the move is accepted, the collective variable of replica \( a \) performs a jump from \( s^a(x^a) \) to \( s^a(x^b) \), and for replica \( b \) from \( s^b(x^b) \) to \( s^b(x^a) \). The exchange moves are not introduced for ensuring convergence to any distribution of states, but to introduce a jump process on top of the standard molecular dynamics evolution, which greatly increases the capability of each replica to diffuse in the CV space. Even if the bias is defined in a few collective variables at a time, each configuration after several accepted exchanges will eventually explore the space spanned by all the collective variables. This greatly improves the capability of the system to explore the configuration space available to it.

It is evident that it is not possible to reconstruct univocally a two-dimensional free energy surface from its one-dimensional projections. However, in this method the unbiased distribution of states can be obtained from an additional replica of the system that is not biased by any time-dependent potential (the neutral replica). This replica is allowed to exchange with the others according to Equation 2.23, but its biasing potential is zero. Introducing this replica has been found useful because when \( \frac{\alpha}{\tau_G} \) is small, it samples approximately the canonical distribution associated with the true (unbiased) potential of the system. In practical applications (namely for finite \( \frac{\alpha}{\tau_G} \)) the neutral replica distribution approaches the canonical distribution only approximately, with an accuracy that is comparable to that of the reconstructed free energies [63].
2.10 Simulation procedures

The general procedure for the construction and the running of the protein simulations of investigation in this thesis is described in this section. The details such as parameters used, cutoffs, box dimensions, specific methodologies implemented, simulation times etc. are described in the Computational Details of each chapter.

For each simulation the system was built and calculations were performed using the simulation package Gromacs 3.3 [209]. The starting structures coordinate of insulin and apoC-II were obtained from the Protein Data Bank (PDB, http://www.rcsb.org/pdb/). The protein under study was initially stripped of all the hydrogens, before it was re-built with the new parameters of a selected force field. After the topology file for the protein was made, a simulation box was constructed with chosen proportions. The dimensions of the box can vary depending on the system’s size, however a minimum distance of 10 Å was set from the periphery of the protein to the edge of the box to eliminate cross-interactions. The simulation box was then filled with explicitly represented water molecules until water density of \( \sim 1.0 \text{ g/cm}^3 \) was obtained. If the solutes had a non-zero total charge, counterions were incorporated in the simulation box to neutralise the system.

Once the construction of the system had been finalised, energy minimisation was performed to correct for bad van der Waals contacts or steric clashes between the protein and water. The first-order minimisation method of steepest descent was applied to reach the lowest local energy point, where minimisation was considered converged when the maximum force exerted on the system was smaller than 20 kJ/mol/nm.

The solvent equilibration stage consisted of constant pressure (NPT) simulation at \( T = 300 \text{ K}, P = 1 \text{ bar} \) with protein atoms restrained to allow the solvent to relax around the solute. Equilibration of the entire system was then performed where the protein was allowed to move. Following the equilibration stages, data collection for analysis was performed at NPT ensemble with constant temperature at 300 K and pressure of 1 bar. The \textit{leap-frog} algorithm is utilised by default in the Gromacs package to integrate Newton’s equations of motion.
Most of analysis of the MD trajectories was performed using the analytical tools of the Gromacs 3.3 package. The secondary structure evolution plots were calculated using the algorithm STRIDE [210]. This method utilises hydrogen bond energy and mainchain dihedral angles in addition to hydrogen bond distances to classify the secondary structure of a protein. Visualisation and structural representation of the systems were performed using Visual Molecular Dynamics (VMD) [211].
Chapter 3

3. Proteins investigated: Insulin & ApoC-II

3.1 Overview

In this thesis computational techniques were utilised to gain insight into the protein behaviour during folding, misfolding and aggregation under various conditions. Specific applications related to two important proteins; insulin and apolipoprotein C-II (apoC-II) were investigated.

Insulin is an excellent candidate for a computational study because of the extensive range of literature available on its structure and function under various conditions. Results of simulation studies of insulin can be easily compared with previous experimental and theoretical findings. ApoC-II is the protein experimentally studied by our group of external collaborators (Assoc. Prof. Geoff Howlett and his team from the Bio21 Institute) thus comparison of our theoretical models with experimental data was possible. Our primary goal was to assist the experimental team with interpretation of their data and verification of the proposed aggregation mechanisms of apoC-II and its derived peptides. Discussion on the structure and function of insulin and apoC-II, as well as some of the more recent studies in their folding and aggregation behaviour is presented in Sections 3.2 and 3.3, respectively of this chapter.
3.2 Insulin

Insulin is a protein hormone secreted by the β cells of the islets of Langerhans in the pancreas. It has the important biological function of regulating the metabolism of carbohydrates and fats, especially the conversion of glucose to glycogen which lowers the blood glucose levels. High sugar levels are a result of reduced secretion or activity of insulin, which can have detrimental effects on the human metabolism. Diabetes mellitus, in particular Type 1 and Type 2 diabetes are forms of pathologic conditions resulting from disruption in insulin secretion or function. Insulin and diabetes-related diseases can be inherited or developed at any stage during life, and these diseases are affecting an increasing number of people throughout the world. Following the discovery of insulin and its involvement in many diseases, the chemical, biological and physiological properties of insulin have been widely investigated. As of April 2009, a search in the PubMed database resulted in 245 620 articles having the word insulin in their title.

Some of the more significant initial discoveries in protein sciences are related to insulin and its structure, for which several Nobel prizes were awarded. Banting and Best [212] are credited with discovering that the administration of pancreatic extracts had the ability to reduce sugar (glucose) levels in the blood, and named their extracts insulin. In 1923 they were awarded the Nobel Prize in Physiology or Medicine for the discovery of insulin. At that time, the patent for insulin was sold to the University of Toronto for one dollar. It was John J. Abel [213] who obtained the first crystals of insulin in its rhombohedral form, although it was David A. Scott [214] who discovered that the rhombohedral crystal obtained was actually a zinc-insulin complex, knowing that the pancreas contains zinc.

3.2.1 Insulin structure

In 1954, molecular biologist Frederick Sanger determined the complete amino acid sequence of the insulin molecule [215]. It was the first protein to have its structure completely identified. He was awarded the Nobel Prize in Chemistry in 1958 for his discovery. This study showed that monomeric insulin is composed of two separate chains, A and B, containing 21 and 30 amino acids, respectively. The residue sequences of chains A and B of porcine insulin are listed in Table 3.1. The monomer contains three disulfide bonds; one is an
intra-A chain disulfide A6-A11 and two inter-AB chain disulfides, A7-B7 and A20-B19, which clamp chain A helices at the end of the central chain B helix. A cartoon representation of the monomer structure and disulfide pairing between chains A and B is shown in Figure 3.1.

### Chain A

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Gly</td>
<td>Ile</td>
<td>Val</td>
<td>Glu</td>
<td>Gln</td>
<td>Cys</td>
<td>Cys</td>
<td>Thr</td>
<td>Ser</td>
<td>Ile</td>
</tr>
<tr>
<td>10</td>
<td>Cys</td>
<td>Ser</td>
<td>Leu</td>
<td>Tyr</td>
<td>Gln</td>
<td>Leu</td>
<td>Gln</td>
<td>Asn</td>
<td>Tyr</td>
<td>Cys</td>
</tr>
<tr>
<td>20</td>
<td>Asn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Chain B

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Phe</td>
<td>Val</td>
<td>Asn</td>
<td>Gln</td>
<td>His</td>
<td>Leu</td>
<td>Cys</td>
<td>Gly</td>
<td>Ser</td>
<td>His</td>
</tr>
<tr>
<td>10</td>
<td>Leu</td>
<td>Val</td>
<td>Gln</td>
<td>Ala</td>
<td>Leu</td>
<td>Tyr</td>
<td>Leu</td>
<td>Val</td>
<td>Cys</td>
<td>Gly</td>
</tr>
<tr>
<td>20</td>
<td>Glu</td>
<td>Arg</td>
<td>Gly</td>
<td>Phe</td>
<td>Phe</td>
<td>Tyr</td>
<td>Thr</td>
<td>Pro</td>
<td>Lys</td>
<td>Ala</td>
</tr>
</tbody>
</table>

Table 3.1 Primary sequence of chain A and B of porcine insulin (PDB code 1ZNI).

Figure 3.1 Representation from different angles of the porcine insulin monomer. Chain A is coloured blue, chain B is coloured red and the disulfide bonds between the Cys residues are coloured yellow.
The formation of the insulin monomer is believed to be mediated by a third chain, chain C, located between the two chains, which is cleaved in the production process, along with several extra residues from the N-terminus of chain B [216]. At micromolar concentrations insulin forms dimers by hydrogen bonding at residues B24 and B26 [217, 218]. These hydrogen bonds link the two C-termini of the two monomers and form anti-parallel \(\beta\)-sheet structure. In the presence of zinc ions insulin constitutes as a hexamer, and in this form it is stored in the pancreas. Upon entrance into the serum, the hexamer dissociates and binds to its receptor as a monomer [216]. The three-dimensional (3D) X-ray crystal structure of the porcine two-zinc insulin hexamer (Figure 1.3, Chapter 1) was determined in 1969 to a resolution of 2.8 Å by Hodgkin and co-workers [219, 220]. Five years prior to her discovery of the structure of insulin, she was awarded the Nobel Prize in Chemistry for her determinations by X-ray techniques of the structures of important biochemical substances such as penicillin and vitamin B12. Subsequent refinements of the crystal structure of insulin have been performed to resolution of 1.9 Å [221], 1.5 Å [222] and 1.2 Å [223]. This final refinement allowed the position of hydrogen atoms to be also determined.

Insulin structure varies slightly between species of animals, and the differences are found mainly at locations A8, A10 and B30 of their residue sequence. The variations in residues between the human, porcine and bovine insulin are listed in Table 3.2. Due to the low variation between these forms of insulin, there is very little antigenicity associated with their use in humans. Their biological activity is also similar, suggesting that the residues which vary between the species are not important in receptor binding, nor do they affect the structure of the hormone. For our studies (presented in Chapter 4 and 5) we used the porcine form of insulin (PDB code 1ZNI), which was extensively investigated by our research group [224-230].

<table>
<thead>
<tr>
<th>Species</th>
<th>Position and Residue</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A8</td>
</tr>
<tr>
<td>Human</td>
<td>Thr</td>
</tr>
<tr>
<td>Porcine</td>
<td>Thr</td>
</tr>
<tr>
<td>Bovine</td>
<td>Ala</td>
</tr>
</tbody>
</table>

Table 3.2 Specific sequence differences between human, porcine and bovine insulin.
There are many structures currently available of insulin, and the general binding mode of the insulin-receptor complex is known [231]. Chain B of insulin is believed to retain much of its structure independently of chain A [232-234], thus it is an excellent candidate for computational investigations. The secondary structure characteristics of chain B of insulin are generally defined as the N-terminus (residues Phe1 to Gly8), a central \(\alpha\)-helix (residues Ser9 to Cys19), a characteristic chain B fold (\(\beta\)-turn, residues Gly20 to Gly23) and the extended C-terminus (residues Phe24 to Ala30). There are several known conformational states for the N-terminus of chain B [105, 235-240]. The principal conformations have been designated as the R-state and T-state [241], although additional conformations have also been identified [242]. A schematic showing several conformational states is presented in Figure 3.2. The T-state is associated with insulin’s activity as it is believed to be the monomeric solution state [243]. In the T-state, chain B consists of an \(\alpha\)-helix (Ser9 to Cys19) and extended N- and C-termini regions [220, 244]. In the R-state, the central \(\alpha\)-helix fold extends into the N-terminal residues, forming \(\alpha\)-helix structure from Phe1 to Cys19 [245, 246]. Another known variation is the “frayed” or Rf-state, where the \(\alpha\)-helix is only present for some of the N-terminal residues (Gln4 to Gly8) in addition to the central helix.

Figure 3.2 An overlay of X-ray structures, showing experimentally observed states of insulin chain B.
3.2.2 Insulin folding and aggregation studies

The structural variation among the crystallographic forms has generated interest in the dynamics and folding of insulin. Keller et al. [247] examined the structure, folding and biological activity of human insulin mutants, Thr(B27)Pro and Pro(B28)Thr insulin (PT insulin), in aqueous solution using NMR spectroscopy. They discovered that in pure water PT insulin has a very flexible structure with biological activity 50% higher than that of the native form. The flexibility and overall tertiary structure of insulin were determined as the key factors for expressing biological activity of insulin. Structure-activity studies of insulin indicate that the C-terminus of chain B is integral to receptor interaction [248-250] and also suggest that the conformation of the C-terminus is influenced by the structure of the N-terminus [235, 238, 251]. Various studies of insulin [236, 252], including a preliminary crystallographic structure of the native insulin monomer at a low pH [253], and also solution structures of isolated chain B determined by NMR spectroscopy [232, 233], confirm both termini’s mobility.

One of the initial computational investigations performed on insulin were those of Kruger et al. where they applied MD simulations to study the independent behaviour of both monomers present in the rhombohedral 2Zn porcine insulin, in the absence of quaternary structure and crystal packing forces. In the 120 ps simulations they discovered partial relaxation in the conformation from the crystallographic form, however, the secondary structure of the molecule was preserved [254]. Longer simulation performed on crystalline 2Zn insulin showed that in the absence of crystal contacts both monomeric and dimeric insulin have a high degree of intrinsic flexibility [255]. In particular, the dynamic behaviour of the dimer was found to be asymmetric [256].

The variation in structure in the N-terminal region of insulin chain B lead several teams to closely investigate the transition pathways between the T and R states of chain B [257, 258]. Using targeted molecular dynamics, Schlitter et al. determined that the transition from R (folded) to T (unfolded) state was more favourable than from the T to R state [259]. The intrinsic flexibility of the N- and C-termini of insulin and the possible role they have in the binding process to the receptor has been computationally investigated by Zoete et al. [260]. Furthermore, in order to determine the residues that are most important for the stability of insulin, the same team performed computational alanine scanning where they found mutations of LeuA16, TyrA19, LeuB11, LeuB15 and ArgB22, may result in misfolding and inactivity of the insulin analogues [261].
Budi et al. performed MD simulations on chain B of insulin under the influence of static and oscillating fields, ranging from $10^7$ to $10^9$ V/m [227]. They found that both variants have an effect on the normal behavior of the protein, with oscillating fields being more disruptive to the structure compared to static fields of similar effective strength. The application of a static field had a stabilising effect on the secondary structure, restricting the inherent flexibility that is crucial for insulin’s biological activity. This inherent flexibility was also observed in a computational study of possible thermal and chemical effects on the dynamics of chain B [225].

When a protein is subjected to external stress (e.g. thermal, electric or chemical) it can experience conformational changes that can cause the protein to misfold, and such structural changes can initiate fibril formation. Insulin is one of the protein family members known to form fibril aggregates. Dissociation of the hexamer into a monomer is required for binding to the receptor and subsequent biological activity [216]. However, under specific conditions such as exposure to acidic pH, elevated temperature, agitation, or contact with hydrophobic surfaces, dissociation of the hexamers and partial unfolding of the monomer is facilitated, which ultimately leads to aggregation and subsequent amyloid fibril formation [262-264]. The properties of insulin, such as, self-assembly, fibrillation, crosslinking and denaturation have created problems in production, storage and use in therapy. In particular, insulin has the tendency to form fibrils in infusion pumps, because it is subjected to agitation and increase in temperature [265]. Amyloid deposits containing intact insulin molecules have been discovered at sites of repeated insulin injections in diabetic patients [266].

A model for the organisation of insulin molecules in the insulin fibril has been proposed by Brange et al. based on the crystal structure of the monomeric insulin mutant, despentapeptide (B26-B30) insulin (DPI) [267]. Two sets of hydrophobic interactions in the fibril model were suggested. One set of contacts involved the aliphatic residues, Ile (A2), Val (A3), Leu (B11), and Leu (B15), which became exposed when the C-terminus of chain B was displaced. These residues came in contact with a hydrophobic surface consisting of Leu (A13), Leu (B6), Ala (B14), Leu (B17), and Val (B18) that are normally buried when three insulin dimers form a hexamer. The second set of contacts in the fibril model was considered to be an antiparallel $\beta$-sheet structure between the residues Phe (B1) and His (B5), running perpendicular to the axis of the fibrils. The $\beta$-sheet was thought to be a favourable contact through which bundles of fibrils may develop from individual fibrils.
To test Brange’s model, the molecular basis of insulin fibril formation was investigated by Nielsen and co-workers where they studied the structural properties and kinetics of 20 different human insulin mutants at both low pH (conditions favouring monomer/dimer) and at pH 7.4 (conditions favouring tetramer/hexamer) [268]. Their results indicated that the monomer is the state from which fibrils arise. The insulin mutants were found to retain native-like secondary and tertiary structures under all conditions studied. Two surfaces on the insulin monomer were identified as potential interacting sites in insulin fibrils, one consisting of the His (B10), Tyr (B16), and Leu (B17) residues and the other consisting of the Thr (A8) and Phe (B25) residues. The importance of both hydrophobic and electrostatic interactions in the initial stages of fibrillation was demonstrated by the marked increase in the lag time of fibril formation for mutations to more polar and charged residues.

In recent atomic resolution crystallographic studies performed in Eisenberg’s laboratory, the $^{11}$LVEALYL$^{17}$ sequence from human insulin chain B was identified as a potentially fibrillogenic region [110]. They investigated nucleated growth by preparing crystalline seeds from the insulin segment LVEALYL, from which they determined the steric-zipper structure of $^{12}$VEALYL$^{17}$. These seeds were found to shorten the characteristic lag time for the growth of insulin fibrils as much as seeds prepared from fibrils of the entire insulin molecules, whereas a non-fibril forming segment of insulin had much longer lag time. The microcrystal structure of the unit-cell and pair of $\beta$-sheets formed by the peptide segment VEALYL of insulin is shown in Figure 3.3. The fibrils formed by VEALYL peptides have parallel $\beta$-sheets, and each $\beta$-sheet is composed of in-register $\beta$-strands arranged in an anti-parallel conformation. This peptide is a very good model for future computational investigations of the fibrillation mechanisms of insulin, because of its small size and the availability of experimental data. Discussion of ideas for continued work in this area is presented in the Future work chapter of this thesis.
In this project, insulin was used as a model for our studies on the possible structural biases different force fields may exert on the conformation and dynamics of chain B. The rich experimental and computational literature available on this protein makes it a perfect candidate for this type of investigation. The results from our systematic comparison of different force fields for computational studies of insulin are presented in Chapter 4. Furthermore, the folding free energy landscape of chain B of insulin was investigated using the BE-META technique. The findings from this study are presented in Chapter 5 of this thesis.

### 3.3 Apolipoprotein C-II

Apolipoprotein C-II (apoC-II) is a 79 residue protein and a member of the very-low-density lipoproteins (VLDL). It plays an important physiological role as an activator of lipoprotein lipase [269]. Lipoprotein lipase is an enzyme found in the capillaries that catalyses the hydrolysis of triacylglycerols in chylomicrons to glycerol and fatty acids [270]. A human deficiency of apoC-II is associated with elevated triacylglycerol levels in the blood, which is consequently linked to atherosclerosis, and by extension, the risk of heart disease and stroke.
3.3.1 ApoC-II structure

ApoC-II binds reversibly to the polar lipid surface of plasma lipoprotein particles in vivo and associates with a range of natural and synthetic lipid surfaces in vitro with an accompanying change in secondary structure. Analysis of this conformational change and the amino acid sequence of the lipid binding regions has led to the generally accepted hypothesis that apolipoproteins associate with lipid surfaces by means of amphipathic helices predicted to be present in all members of the family [271]. An amphipathic helix is defined as an $\alpha$-helix with opposing polar faces oriented along the long axis of the helix.

The structure of apoC-II was initially determined using NMR techniques in the presence of sodium dodecyl sulphate (SDS) micelles by MacRaild et al. [272] and Zdunek et al. [273]. However, the lack of similarity between the SDS micelles and lipoprotein particles made the interpretation of the data problematic, particularly in terms of lipid interactions. Several years later, MacRaild and co-workers applied similar techniques to determine the structure of apoC-II in the presence of dodecylphosphocholine (DPC) micelles, which proved to be a good alternative to SDS micelles [274]. DPC is a detergent that shares the phosphocholine headgroup with the phosphatidylcholine class of biological phospholipids, and by exploiting this similarity they were able to infer details of the apoC-II/lipoprotein interaction and determine the apoC-II lipid bound structure.

The primary sequence of human apoC-II is listed in Table 3.3. The region between residues 1 to 12 is highlighted in grey because the structure of this region was too difficult to assign [273, 274]. The secondary structure of apoC-II is well defined, with $\alpha$-helical elements over most of the sequence, specifically between residues Leu15 to Lys39 and Tyr63 to Gly77. The central region, comprising residues Thr40 to Gly65 is substantially disordered, containing predominantly coil and turn structures. A representation of the secondary structure of apoC-II(13-79) is shown in Figure 3.4.
CHAPTER 3. Proteins investigated: Insulin & ApoC-II

ApoC-II

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Thr</td>
<td>Gln</td>
<td>Gln</td>
<td>Pro</td>
<td>Gln</td>
<td>Gln</td>
<td>Asp</td>
<td>Glu</td>
<td>Met</td>
<td>Pro</td>
</tr>
<tr>
<td>10</td>
<td>Ser</td>
<td>Pro</td>
<td>Thr</td>
<td>Phe</td>
<td>Leu</td>
<td>Thr</td>
<td>Gln</td>
<td>Val</td>
<td>Lys</td>
<td>Glu</td>
</tr>
<tr>
<td>20</td>
<td>Ser</td>
<td>Leu</td>
<td>Ser</td>
<td>Ser</td>
<td>Tyr</td>
<td>Trp</td>
<td>Glu</td>
<td>Ser</td>
<td>Ala</td>
<td>Lys</td>
</tr>
<tr>
<td>30</td>
<td>Thr</td>
<td>Ala</td>
<td>Ala</td>
<td>Gln</td>
<td>Asn</td>
<td>Leu</td>
<td>Tyr</td>
<td>Glu</td>
<td>Lys</td>
<td>Thr</td>
</tr>
<tr>
<td>40</td>
<td>Tyr</td>
<td>Leu</td>
<td>Pro</td>
<td>Ala</td>
<td>Val</td>
<td>Asp</td>
<td>Glu</td>
<td>Lys</td>
<td>Leu</td>
<td>Arg</td>
</tr>
<tr>
<td>50</td>
<td>Asp</td>
<td>Leu</td>
<td>Tyr</td>
<td>Ser</td>
<td>Lys</td>
<td>Ser</td>
<td>Thr</td>
<td>Ala</td>
<td>Ala</td>
<td>Met</td>
</tr>
<tr>
<td>60</td>
<td>Ser</td>
<td>Thr</td>
<td>Tyr</td>
<td>Thr</td>
<td>Gly</td>
<td>Ile</td>
<td>Phe</td>
<td>Thr</td>
<td>Asp</td>
<td>Gln</td>
</tr>
<tr>
<td>70</td>
<td>Val</td>
<td>Leu</td>
<td>Ser</td>
<td>Val</td>
<td>Leu</td>
<td>Lys</td>
<td>Gly</td>
<td>Glu</td>
<td>Glu</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3 Primary sequence of apoC-II (PDB code 1SOH). The sequence coloured in grey represents the residues with undefined structure.

Figure 3.4 Ribbon representation of apoC-II(13-79) showing secondary structure and side-chains.

C terminus

N terminus

69
3.3.2 ApoC-II misfolding and aggregation studies

In the presence of lipids, apoC-II is comprised of mostly α-helical structures, however in lipid-poor conditions apoC-II readily aggregates into twisted ribbon-like fibrils. This protein is one of the few amyloid systems to form fibrils at physiological pH without prolonged agitation. ApoC-II and several other apolipoproteins have been identified in atherosclerotic plaques in vivo[275] and it is found colocalised in human coronary artery plaques with serum amyloid P, a non-fibrillar marker of amyloid deposits [276]. The fibrils formed by apoC-II in vitro are homogeneous when viewed by electron microscopy and display a characteristic X-ray diffraction pattern, indicating the formation of cross-β structure [277].

A unique property of mature apoC-II fibrils is that they are relatively small and remain soluble, allowing characterisation of fibril formation by analytical ultracentrifugation. Binger and co-workers used this technique to show that mature fibrils assemble via a reversible pathway that includes fibril breaking and joining [278]. Solution conditions, such as the presence of auxiliary proteins [279], macro-molecular crowding agents [280] and the presence of phospholipids below and above the critical micelle concentrations (CMC) alter the kinetics of apoC-II fibril formation and fibril morphology [142, 144]. Specifically, micellar phospholipid concentrations were found to inhibit apoC-II fibril formation, whereas sub-micellar concentrations accelerate the fibril forming process. Recent experimental work explored the effect of oxidation and mutation of methionine residues on the amyloidogenic nature of the protein [281]. Oxidation of the Met residues (Met9 and Met60) with hydrogen peroxide inhibited fibril formation. However, the oxidised monomers were found to have no effect on the already formed native apoC-II fibrils. This finding suggested that oxidised molecules have a reduced ability to interact with growing fibrils. Additionally, single Met60Val and Met60Gln mutations were also performed. The Met60Val mutant assembled into fibrils with similar kinetics to the wild-type apolipoprotein. In contrast, the Met60Gln mutant showed a significant delay in fibril formation when compared to the wild-type, however the fibril formation was not completely inhibited. This finding together with the oxidation of Met60 have showed that a change in hydrophobicity at a single position in the apoC-II core region is sufficient to significantly reduce the fibril-forming ability of this protein [281].

Wilson and co-workers performed hydrogen/deuterium exchange and proteolysis studies to identify the core region(s) within apoC-II fibrils that may stimulate amyloid formation [282]. The peptide fragments composed of residues 60 to 70 and 56 to 76 have been
shown to exhibit an inherent propensity for amyloid fibril formation in solution. The synthetic apoC-II(56-76) readily formed fibrils, although with a different morphology and thioflavin T fluorescence yield compared to full-length apoC-II. Furthermore, Wilson et al. postulated that the ability of apoC-II(60-70) to independently form amyloid fibrils may be the underlying cause of apolipoproteins and their peptide derivatives to accumulate in amyloid deposits \textit{in vivo} [282]. The location on the full-length apoC-II protein and residue composition of the amyloidogenic region 60 to 70 is shown in Figure 3.5.

![Figure 3.5 Amyloidogenic peptide ApoC-II(60-70).](image)

The amyloidogenic properties of apoC-II have been extensively studied \textit{in vitro} and these studies form the basis for much of the current knowledge of amyloid fibril formation by apolipoproteins [283]. Recently, \textit{in silico} studies were performed to gain insight into the atomistic detail of the dynamics and association in various environments of apoC-II peptides.

In a recent investigation, experimental and computational techniques were utilised to explore the effect of oxidation and mutation of Met60 to Val and Gln, on the stability of the apoC-II(56-76) peptide [284]. The wild-type apoC-II(56-76) and 56-76Met60Gln peptides readily assembled into fibrils with similar lag phase, whereas the apoC-II(56-76) with oxidised Met formed fibrils with much slower kinetics and was not completely inhibited. The slowest peptide to form fibrils was the 56-76Met60Val, which exhibited totally different aggregation kinetics compared to the full-length apoC-II with the same mutation. Simulations performed by Legge and co-workers revealed that the behaviour of the apoC-II(56-76) peptide system is different to that of the full length protein system, suggesting that the mechanism of fibril formation is also different [284].

71
Environmental conditions, including pH [285] and presence of lipids, as well as the effect of methionine oxidation have been investigated on the amyloidogenic peptide apoC-II(60-70), where distinct differences in secondary structure have been identified. At fibril-favouring conditions (neutral and low pH) the peptide preferentially adopts structures similar in shape to a \( \beta \)-hairpin, while under fibril-disruptive conditions (lipid-rich and oxidised Met) significantly different structures were obtained. The formation of a strong hydrophobic core, as exhibited in a \( \beta \)-hairpin structure, have been suggested favourable for peptide interaction and subsequent fibril formation [285].

In a more recent study, the effect of sub-micellar concentrations of lipids on amyloid fibril formation by the apoC-II(60-70) peptide was investigated using computational and experimental techniques [141]. Thioflavin T fluorescence studies showed that sub-micellar levels of the short-chain phospholipids D5PC and D6PC strongly inhibited amyloid formation by the apoC-II(60-70) peptide. In contrast, the sedimentation equilibrium analysis of these peptide-lipid mixtures indicated the presence of soluble oligomeric complexes. Using the umbrella sampling method, Hung et al. determined the dimerisation free energies of several dimer complexes in the presence and absence of lipids. Their simulations indicated that the short-chained phospholipids promote the formation and stabilisation of dimers by enhancing their inter-subunit hydrophobic interactions. Based on the experimental and computational results Hung et al. proposed that peptide-bound lipids inhibit amyloid fibril formation by trapping of dimers and other oligomeric species in diverse non-fibril forming conformations, reducing their likelihood of acquiring subunit conformations prone to fibril nucleation and growth [141].

In Chapter 6 of this thesis the effect of mutation on the dynamics and structure of the shorter amyloidogenic peptide apoC-II(60-70) was investigated. The effects of different lipid concentrations on the conformation of this peptide were also investigated and the results are presented in Chapter 7. Finally, the structural stability of pre-formed oligomeric species of various sizes and arrangements was investigated. Preliminary results from this study are presented in Chapter 8.
3.4 Project aims

The intent of this project is to apply molecular dynamics techniques and derivative methods to investigate the folding, misfolding and aggregation of peptides. Applications related to two important proteins, insulin and apolipoprotein C-II (apoC-II), are investigated. The specific aims of the project are categorised below:

(1) Determination of the biases and effects of various force fields on the simulated structure and dynamics of insulin. The choice of force field in protein simulations is an important one, as it will ultimately influence the quality of the results.

(2) Investigation of the folding free energy surface of chain B of insulin using the novel Bias-Exchange Metadynamics (BE-META) technique. The identification of the folding mechanisms, pathways and transition states are of particular interest.

(3) Investigation of the propensities for fibril formation in wild-type and mutated (hydrophobic/hydrophilic residues) apoC-II(60-70) peptide monomers.

(4) Investigation of the effects of lipids and their concentration on the structure and dynamics of apoC-II(60-70), in order to identify fibril forming and inhibiting mechanisms and conditions.

(5) Determination of the size, shape and orientation of the nucleus or “seed” for fibril formation of apoC-II(60-70).
Chapter 4

4. Systematic comparison of empirical force fields for molecular dynamic simulation of insulin

4.1 Overview

The use of atomistic simulation methodologies based on empirical force fields has enhanced our understanding of many physical processes governing protein structure and dynamics. However, the force fields used in classical modelling studies are often designed for a particular class of proteins and rely on continuous improvement and validation by comparison of simulations with experimental data.

A comprehensive comparison of five popular force fields for simulation of insulin is presented. The effect of each force field on the conformational evolution and structural properties of the protein is analysed in detail and compared with available experimental data. In this study we observed that different force fields favour different structural trends. We found that the all-atom force field CHARMM27 and the united-atom force field GROMOS 43A1 delivered the best representation of the experimentally observed dynamic behaviour of chain B of insulin.
CHAPTER 4. Systematic comparison of empirical force fields for molecular dynamics simulations of insulin

4.2 Introduction

In the last few decades molecular dynamics (MD) simulations have emerged as a powerful tool for the characterisation of biomolecular structure and dynamics at the atomic level. The technique has helped us understand complex molecular processes associated with protein conformational changes, ranging from studies of enzyme-reaction mechanisms and ligand binding to problems of protein re-folding and denaturation (see references within Ponder and Case [286]). Fundamental to such simulations are the forces that govern the atomic motions, derived from a pairwise atom-atom interaction function referred to as an empirical potential energy function or a force field. Details on force field theory and development are presented in Chapter 2. The choice of mathematical function describing a force field is important, since it will ultimately determine the quality of the results. In order to understand possible systematic effects of force fields on the simulated dynamics of insulin a detailed comparison of five commonly used force fields was performed.

The biomolecular force fields selected for our study include the recent versions of the all-atom CHARMM27 [150], AMBER03 [165], OPLS [163] and the united-atom GROMOS 43A1 [174] and GROMOS 53A6 [153] force fields. The parameters for all five force fields were extensively optimised over a number of years with particular emphasis on the treatment of proteins. Detailed discussion on the parameterisation and functions of each force field is presented in Chapter 2. AMBER03 was selected for this study, because it showed very good agreement with experimental data in an extensive force field investigation between the readily available AMBER force fields [287], not considering some custom modified AMBER versions.

An important consideration in the use of a force field for biomolecular simulations is the accurate treatment of the condensed aqueous environment. When selecting a water model to use for a particular study, its compatibility with the biomolecular force field used is of crucial importance. The reason for this is that most force fields have been developed in conjunction with a specific water model. For example, AMBER, CHARMM and OPLS have been developed with the TIP3P water model, OPLS also with TIP4P while GROMOS with SPC and F3C models.
Recent simulations of three proteins performed using the AMBER94, CHARMM and OPLS force fields, concluded that these force fields behave comparably or at least that it was not possible to distinguish between the force fields in the 2 ns time scale investigated [288]. In a more recent article by Villa et al. the sensitivity of molecular dynamics simulation to different force fields was examined [289]. The three parameter sets (43A1, 53A5 and 53A6) of the GROMOS force field were used to simulate 36 structures of 31 proteins. In their investigation, no major differences were detected in a wide range of structural properties such as the root-mean-square deviation from the experimental structure, radii of gyration, solvent accessible surface, secondary structure, or hydrogen bond propensities on a 5 to 10 ns time scale, despite the differences in the force field parameters. On the contrary, Mu et al. showed that the recent version of the AMBER, CHARMM, GROMOS, and OPLS force fields used in their study of trialanine differ considerably in the description of the dynamical properties of the system. It was found that the lifetime of some conformational states differ by more than an order of magnitude, depending on which force field model was used [169]. Several empirical force fields and quantum mechanics/molecular mechanics (QM/MM) force fields were used by Hu et al. to investigate the sampled conformations of unfolded polypeptide chains in aqueous solution [170]. This group also found variation in their results which they related to the different force fields that were applied, rather than to a direct comparison with experimental data. The findings by Sorin and Pande [290] on helix-coil systems in conjunction with work by Zeman et al. on alanine dimers and trimers [291] suggested that a dependence of force field accuracy on peptide length exists even for simple models such as polyalanine-based systems. In the article by Yoda et al. [292] discrepancies in simulation results due to the choice of force field have been well documented for a significant number of systems.

In order to understand any possible systematic bias of the common force fields the protein insulin for the present investigation was selected. Insulin is a widely studied protein and there is an extensive experimental and computational literature available for comparison (summarised in Chapter 3). Molecular dynamics enables sampling of conformational states of a protein under controlled conditions and therefore delivers detailed understanding of the protein’s dynamics. Previous MD simulations using the GROMOS 37c force field, show conformational flexibility of insulin chain B, reporting a high degree of movement in aqueous solution of both the monomer and dimer [256]. Simulations of monomeric insulin in the T-
CHAPTER 4. Systematic comparison of empirical force fields for molecular dynamics simulations of insulin

state conformation were performed by Zoete et al. [260] for 5-10 ns using the CHARMM22 force field where they identified the flexibility of the N- and C-terminal regions of chain B. This inherent flexibility was also observed in a previous computational study of possible thermal and chemical effects on the dynamics of chain B [225] using the NAMD simulation package and the CHARMM27 parameter set. The most recent of these studies performed by Legge and co-workers yielded information clarifying uncertainties about the structure and dynamics of insulin with respect to its biological behaviour by performing multiple MD simulations with the CHARMM27 force field in explicit solvent [230], as an alternative way to improve the conformational sampling.

In this chapter we attempt to identify and compare possible systematic effects of multiple force fields on the structure and dynamics of chain B of insulin in a course of identical molecular dynamics simulations. We also aim to investigate which force field gives the best representation of the experimentally observed conformational features and behaviour of chain B within our selected timeframe. The assessment of the force fields has been performed by comparison of results with experimental data which illustrates the 3-dimensional morphology of the protein, such as X-ray crystallographic structures [241, 242, 245] (see Figure 3.2), Nuclear Overhauser Enhancement (NOE) distance restraints from the solution NMR structure of isolated chain B of insulin [232], and a previous molecular dynamics study of chain B which reproduced known experimental structural features [230]. To the best of our knowledge this is the only information available in the literature for direct comparison of conformational features observed in the simulations with different force fields. Dynamic and structural properties such as the secondary structure evolution, solvent accessible surface area (SASA), radius of gyration, interproton distance violations, have been obtained from each force field simulation for comparison and are presented in the Results and Discussions section of this chapter.

4.3 Computational Details

This study utilises a classical molecular dynamics algorithm [293] as implemented in the GROMACS 3.3 [209] simulation package. To investigate the effect of different potential energy functions on insulin structure and dynamics, the widely used all-atom CHARMM27,
OPLS, AMBER03 and the united-atom GROMOS 43A1 and GROMOS 53A6 force fields were selected. In this chapter they will be referred to as, CHARMM, OPLS, AMBER, GROMOSA1 and GROMOSA6, respectively. For each force field two separate simulations were performed, for 50 ns each, starting from the same initial structure. All the simulations are labelled sequentially, based on the force field used and the run number, such as, for the calculations using AMBER, the simulations are labelled a1 and a2, with CHARMM, b1 and b2, with OPLS, c1 and c2, with GROMOSA1, d1 and d2 and with GROMOSA6, e1 and e2. Each system was immersed in explicit solvent (water), with the water model selected according to that chosen by the developers at the time of parameterisation. Specifically, the TIP3P water model was used for the CHARMM, AMBER and OPLS force fields, while the SPC model was employed for the united-atom GROMOS force fields.

The starting structure was the Rf conformational state of insulin chain B obtained from the crystal structure of porcine insulin (PDB code 1ZNI [245]). The His residues were protonated, and both termini regions and titratable sidechains were charged, resulting in a neutral to acidic pH environment. The protein was enclosed in a periodic box of 60 Å × 60 Å × 60 Å size, then solvated with ~7075 TIP3P or SPC water molecules (depending on the force field chosen), corresponding to water density of ~1.0 g/cm³. The Particle Mesh Ewald [294] (PME) summation was applied to treat long range electrostatic interactions. Atom based cutoff of 12 Å with switching at 10 Å was used for non-bonded van der Waals interactions. All bond lengths were constrained to their equilibrium value using the SHAKE algorithm [192].

The whole system was energy minimised to remove steric clashes using the steepest descent algorithm. Two equilibration stages were performed, each lasting 600 ps. First, equilibration was performed in the constant volume (NVT) ensemble, and then followed by the constant pressure (NPT) ensemble simulation at 1 bar. Constant pressure was achieved by coupling the system to a Parrinello-Rahman barostat [191] with a relaxation time of 4 ps. Constant temperature of 300 K was maintained using the Berendsen thermostat [188]. The data collection stage was performed under constant pressure (NPT) ensemble for 50 ns for each force field simulation accumulated with a 2 fs timestep.

A recent study by Schulten et al. highlighted difficulties in obtaining sufficient sampling for studying folding using classical MD, and the possible effects a force field might have on obtaining the correct folded structure [295]. True thermodynamic equilibrium is not
feasible from simulations such as these; however, the adequacy of equilibration can be demonstrated by monitoring statistical deviations of the properties of the system. The equilibrium was monitored by calculating deviations in potential energy from the mean value for each simulation. During the data collection (equilibrium), no systematic drifts or fluctuations higher than 3% were observed, signifying that the system was properly equilibrated.

4.4 Results and Discussions

4.4.1 Secondary structure evolution
The secondary structure dynamics reveal details of the conformational changes experienced by the protein over the simulation period. The evolution of the secondary structure of insulin chain B as a function of time for each force field system is presented in Figure 4.1. The secondary structure was classified by VMD using the STRIDE algorithm, which utilises hydrogen bond energy and mainchain dihedral angles in addition to hydrogen bond distances [210]. To identify clusters of similar structures in each trajectory, the method described by Daura et al. [296] was implemented, where the RMSD of atom positions between all pairs of structures was determined. For each structure, the number of other structures with RMSD of 0.25 nm or less for the Cα atoms was calculated (neighbours). The structure with the highest number of neighbours was taken as the centre of a cluster (reference structure) and together with all its neighbouring conformations formed a (first) cluster. The members of this cluster were then removed from the pool of structures. The process was then repeated until the pool of structures was empty, resulting in a series of nonoverlapping clusters obtained for each simulation. Cluster analysis was performed on each trajectory using 2500 structures extracted at time intervals of 0.025 ns. The reference structure of the two most populated clusters of each simulation superimposed over the α-helix region (residues Ser9 to Cys19), and colour coded by their residue type can be seen in Figure 4.2.
Figure 4.1 Secondary structure evolution of insulin chain B. Simulations for each force field are labelled as described in the Computational Details section of this chapter. The secondary structure colour codes: magenta – α-helix, red – π-helix, cyan – turn, white – coil, yellow – extended conformation and green – bridge.
Figure 4.2 The reference structure of the two most populated clusters of each force field simulation, labelled as described in the Computational Details section of this chapter. The colouring of each protein is by its residue type, blue – basic, red – acidic, green – polar and silver – nonpolar. The hydrogen bonds of structure d2 are coloured purple.
A variation of secondary structures featured in different force field simulations was found, however, some similar trends can be observed. All systems showed an early change from the starting structure at residue Cys19, most evident in the CHARMM (Figure 4.1b) and GROMOSA1 (Figure 4.1d) simulations. This change can be expected due to the absence of the disulfide bonds with chain A and the associated destabilisation around this residue. Some helical content was maintained in each simulation although with great variation, as the close inspection of Figure 4.1 demonstrates.

Specifically, the AMBER force field simulations preserved the most helical content, the helix ranging from Asn3 to Cys19, maintaining the majority of structural elements of the initial conformation (Figure 4.1a). Some extension of the helix was seen after 11 ns in simulation a1, between Glu13 and Val18, however within the next 13 ns the helix was restored fully and remained stable for the rest of the simulation, demonstrating the AMBER force field tendency to favour the helical conformation [287]. The turn region between residues Cys19 and Arg22 was largely maintained throughout the simulation. Interestingly, this type of secondary structure was observed in a modelling study using CHARMM27 force field [227] where chain B was simulated in a static electric field of $10^8$ V/m, which stabilised the helix by aligning the peptide along the field direction. AMBER’s tendency to overstabilise helical conformers has been observed in other studies [287]. The two most populated clusters in our simulations (Figure 4.2a) confirm this preference, showing very little variance in the helix with some mobility in the C-terminus. It should be noted that the flexibility of the carboxyl terminus of chain B was proposed to be an important feature in receptor binding [237, 297, 298] of insulin and should be featured to deem a simulation reliable.

In comparison to AMBER, CHARMM produced more structural changes within our simulation time (Figure 4.1b). A break in the helix at the Gly8 residue has been seen within the first 5 ns of the two independent calculations, followed by an attempt to retain the helical conformation between residues Gln4 and Cys7. These results are consistent with experimental studies showing that Gly residues are often associated with perturbations in structure and are known as “helix-breakers” [299]. The break in the helix contributed to an increased flexibility of the N-terminus, which enabled the N- and C-termini to move close together, and produced hydrogen-bonded bridges between residues Val2 and Tyr26. This feature can also be seen in the cluster structures of system b1 (Figure 4.2b), where the turn region between residues
Cys19 and Arg22 transformed into a $\pi$-helix after 12 ns and a stable $\pi$-helical conformation was maintained for the rest of the simulation.

It has previously been reported that the formation of $\pi$-helix may be an artifact of a specific force field applied to an MD simulation with implicit solvent [300]. While an implicit solvent simulation cannot be directly compared to simulations in explicit solvent such as those reported here, we must stress that $\pi$-helix has been suggested to be ten times more prevalent than previously believed [301]. Transitions between $\alpha$- and $\pi$-helix were previously reported in MD simulations and are believed to be genuine physical phenomenon and represent an important stage in the helix-to-coil transition [302]. Interestingly, the formation of $\pi$-helices was also observed in the GROMOSA6 force field simulations (Figure 4.1e). In contrast to system b1, there was very little $\pi$-helix formation seen in system b2, where most of the region between residues Cys19 and Arg22, retained its turn conformation.

Noticeable structural changes were seen in both simulations using OPLS (Figure 4.1c), starting with a break at Ser9, which continued into dissipation of the helix. Most helical content was lost by the end of our simulation time, with the majority of the peptide structure adopting a $\beta$-turn conformation. The loss in helical content has been observed in the cluster structures of both simulations using OPLS shown in Figure 4.2c, particularly in the second system c2, where very small amount of helical elements can be seen.

An interesting structural feature observed in the GROMOSA1 systems (Figure 4.1d) was the break in the helix at Gly8, which resulted in an attempt of helix reformation between residues Gln4 and Cys7. Evidence of helical conformation between these residues was observed previously in the NMR solution structure of an insulin chain B mutant [233] and in previous computational study using the CHARMM27 force field [230]. In the present study, the helical region between residues Ser9 and Val18 was well conserved throughout both simulations using GROMOSA1, with some fraying at residue Val18, near the end of simulation d2 (Figure 4.1d). This was likely caused by the hydrogen bonding between the two termini, reflected by the bridge conformation observed in the secondary structure evolution plot of this system. This feature can be seen in the cluster structures of system d2 in Figure 4.2, presented as small parallel $\beta$-sheet with the hydrogen bonds coloured in purple.

The GROMOSA6 force field produced many changes in the secondary structure of the protein from its initial state (Figure 4.1e). A break in the helix at residue His10 occurred during the equilibration stage, which destabilised the protein and lead to a complete loss of
the helical conformation. A transition from $\alpha$-helix to $\pi$-helix to turn conformation can also be seen at the beginning of the simulations. The formation of extended conformation between residues Ser9, His10 and Tyr26, Thr27, represented in yellow in system e2 of Figure 4.2 is associated with the loss of the conserved helical region observed in most of the previous simulations, however no hydrogen bonds were detected. The total loss of helical content can be clearly seen in the most populated clusters of both simulations with GROMOSA6 shown in Figure 4.2e.

### 4.4.2 NOE energies and violations

When assessing the quality of force field, or in the interpretation of ambiguous experimental results, simulated molecular dynamics trajectories of proteins and nucleic acids are often compared with nuclear magnetic resonance (NMR) data [303]. The atom-atom distances derived from the NOE intensities are particularly useful, as they are integral in NMR-based protein structure determination [303] and identification of specific conformational features. The NOE interproton distances are the only experimental data currently available for the description of the 3D conformational features being investigated in this study. The data derived from the NMR structure of bovine insulin chain B in solution [232] provide a useful benchmark for evaluating the conformations observed during the MD calculations. Hawkins et al. derived NOE constraints from a 250 ms NOESY spectrum of the oxidised chain B at 500 MHz, 300 K and pH of 2.2 to 2.5 [232]. As coordinates for the structure were not available, NOE distance restraints were used for comparison.

Using the square-well potential function [304] implemented in the program XPLOR [305, 306], the interproton distance restraints were used to calculate the NOE energy ($E_{\text{NOE}}$) values for the MD conformations from the trajectories of each force field system. The square-well potential function is defined as:

$$E_{\text{NOE}} = \begin{cases} (r_{ij} - r_{ij}^{\text{upper}})^2, & \text{if } r_{ij} > r_{ij}^{\text{upper}} \\ 0, & \text{if } r_{ij}^{\text{lower}} \leq r_{ij} \leq r_{ij}^{\text{upper}} \\ (r_{ij} - r_{ij}^{\text{lower}})^2, & \text{if } r_{ij} < r_{ij}^{\text{lower}} \end{cases} \quad (4.1)$$
CHAPTER 4. Systematic comparison of empirical force fields for molecular dynamics simulations of insulin

where \( r_{ij}^{\text{lower}} \) and \( r_{ij}^{\text{upper}} \) are the lower and upper limit of the target distance restraint, respectively. The distance between a selected set of atoms is averaged according to

\[
( < r_{ij}^{-6} > )^{1/6}
\]  

(4.2)

where \( r_{ij} \) runs through all possible combinations of distance restraints between atoms “i” and atoms “j“ of the selected molecule.

The NOE energy is widely used by experimentalists as a distance restraining function in addition to the standard empirical potential (Equation 2.3, Chapter 2) for simulated annealing, restrained minimisation and molecular dynamics simulations. This energy is a crucial inclusion in various methodologies used for NMR structure determination and refinement (see references within Schwieters and co-workers [307]). In this study the NOE energy was used for evaluation of the conformity of the structures sampled by MD simulations and the NMR distance restraints. To evaluate the results, these pseudo energy values were compared with the \( E_{\text{NOE}} \) of the original PDB structure of each force field system, presented in Figure 4.3. The total number of interproton distance violations, along with the number of “severe” violations (1 Å above the upper bound value of the NOE restraint) is presented in Table 4.1. Several typical structures were selected from the force field simulations and are shown in Figure 4.4, highlighting their satisfied and violated distances, along with the region of their occurrence.

The NOE energies calculated from the simulations using AMBER were mostly higher than the energy of the starting structure (Figure 4.3a). A decrease in energy is seen after 10 ns for system a1 indicating favourable conformational changes. This region of the secondary structure plot corresponds to the turn conformation forming near the \( \beta \)-turn of the peptide. The conformations sampled at \(~12.8\) ns and their NOE distance violations were closely analysed for both simulations using AMBER (Figure 4.4a1, a2) because of their significant NOE energy variance. The structural differences in the turn region of both structures are evident, with the extension of turn conformation observed in simulation a1 and the fully retained helical region in simulation a2. These features produced a large difference in the NOE distances between residues Arg22 and Tyr26 of both structures. In particular, that distance in structure a2 was more than double \(~12\) Å the upper bound value of 6 Å of that restraint, compared to \(~8\) Å in structure a1. The following increase in energy is evoked by the rebuilding of the helix. The turn conformation extension at the C-terminus produced a
CHAPTER 4. Systematic comparison of empirical force fields for molecular dynamics simulations of insulin

decrease in energy near the end of the simulation. Interestingly, very similar NOE violations were observed in a folding study of chain B of insulin using the AMBER03 force field [66] where ~1 µs of data was accumulated using the Bias-Exchange Metadynamics technique [63]. This finding further highlights that adequate sampling has been acquired for this study. Overall, the protein behaviour produced by the AMBER force field, such as the restrictions of atomic movement and the tendency to keep helices well conserved, is not in a good agreement with the NOE restraints of chain B of insulin.

<table>
<thead>
<tr>
<th>Force field</th>
<th>System</th>
<th>Total Violations (309)</th>
<th>Strong 2.7 Å (24)</th>
<th>Medium 3.5 Å (141)</th>
<th>Weak 5 Å (88)</th>
<th>Very weak 6 Å (56)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMBER03</td>
<td>a1</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>a2</td>
<td>9</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>CHARM27</td>
<td>b1</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>b2</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>OPLS-AA</td>
<td>c1</td>
<td>17</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>c2</td>
<td>16</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>GROMOS 43A1</td>
<td>d1</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>d2</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>GROMOS 53A6</td>
<td>e1</td>
<td>14</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>e2</td>
<td>16</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 4.1 Average NOE interproton distance violations calculated for each force field simulation. The total average number of violations, as well as the severe violations larger than 1 Å from the upper bound value of the classified cross-peak intensities by Hawkins et al. [232] are indicated.

In the CHARMM simulations there is a variance in the NOE energy between the two modelled systems (Figure 4.3b), reflected by the structural differences observed. Initially, the NOE energy is below that of the minimised original structure in both simulations, however an increase is seen in system b1 after ~10 ns, where the formation of a π-helix and a bridge between residues Phe24 and Pro28 is observed. This indicates a less favourable secondary structure, which may be due to the bridge formation near the C-terminus. A decrease in energy is seen near the end of the simulation as the bridge disappears, the C-terminus extends and the π-helix reforms. Overall, system b2 produced the lowest NOE energies, indicating that the turn conformation of the chain at residues Leu17 to Cys19 can be more energetically favourable than a π-helix. This finding is in agreement with the distance violations experienced by the structures sampled at ~30.2 ns of the two simulations using CHARMM shown in Figure 4.4. It is evident that the π-helix formation in structure b1 produced more
severe interproton distance violations than the extended turn conformation of structure b2. Turn-like structures were also apparent over residues Leu17 to Glu21 of chain B at the carboxyl terminal in the solution state NMR study performed by Hawkins and co-workers [232].

The simulations using OPLS force field produced the least favourable results, with most of the energies above the NOE energy value of the initial structure (Figure 4.3c). These results correlate well with the calculated interproton violations (Table 4.1), which show that the reduction of helical content is the main contributor to the largest number of violations. The stretching of the protein structure and the violations experienced can be visualised in the conformation sampled at ~20 ns of simulation c1 (Figure 4.4). The helix in chain B of insulin is a well conserved feature observed experimentally while in the OPLS calculations most helical content is lost.

The simulations using the united-atom GROMOSA1 force field agreed with the experimental data better than all the other simulations (Figure 4.3d). In both simulations the NOE energies are below the energy of the initial structure, also having the lowest number of interproton distance violations (Table 4.1). The conformational changes such as the Gly8 helix break are in agreement with the NOE restraints, showing a decrease in energy. This was followed by the formation of a T-like state (associated with insulin’s activity [308]), represented by structure sampled at ~10.24 ns of simulation d1 (Figure 4.4). The small magnitude of NOE distance violations indicates favourable structure. The increase in energy of system d2, near the end of the simulation can be correlated with the close interaction of the termini, however this is probably not an unfavourable event because the NOE energies are still well below that of the starting structure.
Figure 4.3 NOE energy derived from the distance restraints of the sampled structures. The region below the line at \(-41.5 \times 10^3\) a.u. represents the MD structures with NOE energy less than that of the starting structure. The plot shows a moving average of 250 ps.
**Figure 4.4** Important protein conformations and their NOE distance violations. The satisfied distance restraints are shown in yellow and the violated distances in black. The colour scheme is the same as in Figure 4.1.
In contrast to GROMOSA1, GROMOSA6 produced less favourable results (Figure 4.3e). The increase in conformational changes produced large fluctuation in energy with most of the values being above that of the minimised structure. A high NOE energy conformation sampled at ~40 ns of simulation e1 is presented in Figure 4.4, where the extent of the structural deformation and large number of violated distance restraints are clearly seen. As observed with OPLS, the dissipation of the helix is not energetically favoured in the chain B structure.

In summary, the calculated NOE energies allowed us to compare the change in conformation of chain B relative to the starting structure with respect to the NOE intensities during each simulation. The widely used averaging method described by Zagrovic et al. [303] was implemented to give the total violations within each force field system. The highest number of violations were found in the OPLS (17 and 16), and GROMOSA6 (14 and 16) simulations, and the least in the GROMOSA1 (7 and 8). Most violations were found for the weak intensities (Table 4.1) where the same NOEs found in the C-terminal region of the protein were violated in each simulation. All simulations produced a range of structures, however the number of severe violations compared to the total number of restraints, 309, was quite small. This seems to indicate that the NOE restraints allow some degree of flexibility in the molecule, which is in agreement with the range of experimental structures observed [232, 308]. These findings highlight the usefulness of performing multiple simulations in cases like this to explore the range of bioavailable conformations.
4.4.3 Root mean square deviation (RMSD)

The central helix region plays a key role in insulin’s activity. To investigate the effect of each force field on the stability of the helix, RMSD calculations have been performed on this region. The central helix residues (Ser9 and Cys19) in each simulation were aligned and compared to the minimised PDB structure. The RMSD of the central helix for the ten systems is shown in Figure 4.5. The plots show a moving average of 250 ps to enable effective comparison of the overall trends between the systems.

In the AMBER and GROMOSA1 systems, most of the helical region retained its structure throughout the simulation, as illustrated by the most populated cluster structures shown in Figure 4.2. A small increase in RMSD was seen at the end of the simulation d2 of GROMOSA1, generated by the formation of turn conformation at the end of the helix. Transformation of part of the α-helical region into a π-helix (system b1) or turn conformation (system b2) contributed to the sharp increase in RMSD for the CHARMM simulations. However, this was stabilised for the rest of the simulations by the retained helix.

The slow dissipation of the helix observed in the OPLS and GROMOSA6 simulations was obvious from the continuous increase in RMSD. This is almost double the value of RMSD obtained in our previous simulation of thermally stressed chain B at 400 K [230] and points out the limitations of the force fields in reproducing the experimentally observed behaviour of the protein.
Figure 4.5 RMSD of the helical region of chain B, from residues Ser9 to Cys19, compared to the starting structure as a function of time for all systems. The plot shows a moving average of 250 ps.
4.4.4 Radius of gyration

The radius of gyration (Rg) was calculated for each force field system to give an indication of the distribution of atoms relative to the protein’s centre of mass. It is used to measure the change in shape of the protein by comparing it to its starting structure. The average Rg for each force field system, together with its standard deviation, is presented in Table 4.2.

<table>
<thead>
<tr>
<th>Force field</th>
<th>System</th>
<th>Rg (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMBER03</td>
<td>a1</td>
<td>1.026 ± 0.027</td>
</tr>
<tr>
<td></td>
<td>a2</td>
<td>1.007 ± 0.042</td>
</tr>
<tr>
<td>CHARMM27</td>
<td>b1</td>
<td>0.957 ± 0.059</td>
</tr>
<tr>
<td></td>
<td>b2</td>
<td>0.957 ± 0.049</td>
</tr>
<tr>
<td>OPLS-AA</td>
<td>c1</td>
<td>0.976 ± 0.047</td>
</tr>
<tr>
<td></td>
<td>c2</td>
<td>1.004 ± 0.045</td>
</tr>
<tr>
<td>GROMOS 43A1</td>
<td>d1</td>
<td>0.897 ± 0.049</td>
</tr>
<tr>
<td></td>
<td>d2</td>
<td>0.891 ± 0.051</td>
</tr>
<tr>
<td>GROMOS 53A6</td>
<td>e1</td>
<td>1.008 ± 0.055</td>
</tr>
<tr>
<td></td>
<td>e2</td>
<td>1.006 ± 0.061</td>
</tr>
</tbody>
</table>

Table 4.2 Average radius of gyration (Rg) and its standard deviation, calculated for each simulation.

The lowest average radius of gyration was produced by the GROMOSA1 force field, giving a value of 0.897 nm and 0.891 nm for the two trajectories, respectively. These results indicate that the structures formed are closely packed, as can be seen from the most populated cluster structures displayed in Figure 4.2. In addition, the secondary structure evolution plots show continuous interaction between the termini, producing a closed loop of the protein which results in a compact structure. Similar results were produced by the CHARMM force field, though with a slightly higher average Rg, likely caused by some loss of the helical content in the protein.

The highest value of the radius of gyration was produced by the AMBER force field, closely followed by GROMOSA6 and OPLS. For AMBER this is not surprising because of the stability of the inherently elongated helical secondary structure observed in our simulations. For the other two force fields the relatively high Rg indicates the dis ordering of the structure, including the loss of helical content and the lack of specific termini binding of the C-terminus with the helix, contributed to the higher average radius or gyration.
4.4.5 Solvent accessible surface area (SASA)

The SASA was calculated for each force field system using a probe radius of 1.4 Å and Lennard-Jones hard-shell radii for each atom to define the surface. The average SASA of the hydrophobic and hydrophilic residues of chain B, along with the total area, including their standard deviations is presented in Table 4.3.

<table>
<thead>
<tr>
<th>Force field</th>
<th>System</th>
<th>Hydrophobic (nm$^2$)</th>
<th>Hydrophilic (nm$^2$)</th>
<th>Total (nm$^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMBER03</td>
<td>a1</td>
<td>17.47 ± 0.75</td>
<td>11.05 ± 0.54</td>
<td>28.52 ± 1.04</td>
</tr>
<tr>
<td></td>
<td>a2</td>
<td>17.33 ± 0.87</td>
<td>10.81 ± 0.66</td>
<td>28.13 ± 1.35</td>
</tr>
<tr>
<td>CHARMM27</td>
<td>b1</td>
<td>10.88 ± 0.77</td>
<td>17.75 ± 0.86</td>
<td>28.63 ± 1.38</td>
</tr>
<tr>
<td></td>
<td>b2</td>
<td>11.28 ± 0.76</td>
<td>18.29 ± 0.88</td>
<td>29.57 ± 1.37</td>
</tr>
<tr>
<td>OPLS-AA</td>
<td>c1</td>
<td>16.14 ± 0.70</td>
<td>12.10 ± 0.59</td>
<td>28.24 ± 0.99</td>
</tr>
<tr>
<td></td>
<td>c2</td>
<td>16.21 ± 0.87</td>
<td>12.41 ± 0.65</td>
<td>28.61 ± 1.30</td>
</tr>
<tr>
<td>GROMOS 43A1</td>
<td>d1</td>
<td>15.80 ± 1.01</td>
<td>8.96 ± 0.56</td>
<td>24.76 ± 1.26</td>
</tr>
<tr>
<td></td>
<td>d2</td>
<td>16.27 ± 0.98</td>
<td>8.70 ± 0.66</td>
<td>24.98 ± 1.28</td>
</tr>
<tr>
<td>GROMOS 53A6</td>
<td>e1</td>
<td>17.69 ± 0.80</td>
<td>11.04 ± 0.63</td>
<td>28.73 ± 1.06</td>
</tr>
<tr>
<td></td>
<td>e2</td>
<td>17.24 ± 0.84</td>
<td>11.15 ± 0.72</td>
<td>28.38 ± 1.07</td>
</tr>
</tbody>
</table>

Table 4.3 Average solvent accessible surface area (SASA) and standard deviations calculated for each simulation.

The lowest SASA was produced by the GROMOSA1 force field, ~4 nm$^2$ smaller than other systems. This result correlated well with the smallest radius of gyration also observed by this force field, confirming the compactness of the structures explored. The close interaction between the termini which reduced the accessible area to the solvent can be suggested as one of the reasons for the low value. The other force field systems produced very similar values of total SASA. The only significant difference was observed for the hydrophobic/hydrophilic SASA obtained from the CHARMM simulations. Specifically, the SASA for the hydrophobic residues is ~6 nm$^2$ lower compared to the corresponding structures produced by the other force fields, while the exposed area of the hydrophilic residues is ~6 nm$^2$ higher. This result is produced mainly due to the compactness of the structure and the orientation of the side chains. This is in agreement with the accepted theory that proteins will fold to minimise the hydrophobic surface that is exposed to water [309]. As can be seen from the cluster analysis of the CHARMM simulations (Figure 4.2b), the most populated structures are “closed” structures, where a strong interaction between the N- and C-terminus lead to their close proximity, thus reducing the amount of solvent accessing the hydrophobic residues. The
outward orientation of some of the hydrophilic residues, such as Asn3, His5, Ser9 and Lys29, contribute to the slightly larger hydrophilic SASA.

4.4.6 Cross-force field simulations

To illustrate possible systematic bias of the force fields towards particular elements of secondary structure, cross-force field simulations were performed on two predominant structures identified by different force fields. First, an unfolded structure was taken from a GROMOSA6 simulation (Figure 4.2e1) and was simulated using the AMBER03 force field for 15 ns. During this time, the secondary structure of the protein remained unfolded without significant structural change, and no helical formation was observed (Figure 4.6a). This finding, together with the two previously described simulations performed on the native structure of chain B, suggests that the AMBER03 force field may produce a very steep and rugged free energy surface, where a protein can get easily trapped in a local minimum and even a long simulation may fail to sample the remaining conformational space.

Figure 4.6 Secondary structure evolution of insulin chain B; a) Starting structure taken from the last frame of simulation using GROMOSA6 force field was simulated using AMBER03; b) Starting structure taken from the last frame of simulation using CHARMM force field was simulated using OPLS.
CHAPTER 4. Systematic comparison of empirical force fields for molecular dynamics simulations of insulin

Second, one of the most frequently sampled structure by the CHARMM force field (Figure 4.2b2) was simulated using OPLS for 15 ns. Significant change in the conformation of the protein was seen after 7.5 ns of MD, where reduction in helical content became most evident (Figure 4.6b). This result confirms possible favouring of the turn conformation by the OPLS force field as it is in agreement with the behaviour observed in the two simulations of natively folded chain B, described above.

4.5 Conclusions

A systematic comparison of multiple simulations of insulin chain B using five different force fields was performed to gain an improved understanding of the force field influences on the representation of the conformational behaviour of the protein. The effects of these widely used force fields on the secondary structure of insulin and its dynamics were investigated in detail by calculating the conformational evolution, solvent accessible surface area, radius of gyration and interproton distance violations for each force field simulation. Comparisons of our results with X-ray crystallographic structures and NOE distance restraints were undertaken as, to the best of our knowledge, this is the only experimental data available for the description of conformational features we are interested to identify. We have observed that different force fields favour different conformational trends, which is important to be aware of for any classical MD simulation studies.

The all-atom force field CHARMM27 and the united-atom force field GROMOS 43A1 gave the best representation of the experimentally observed dynamic behaviour of chain B of insulin. Similar structural trends were observed with both force fields, including breaking of the helix at Gly8, and the formation of some biologically important conformational states, such as the T and Rf states. The majority of the simulated structures of these two force fields satisfied the NMR distance restraints, although GROMOS 43A1 had lower NOE energy and fewer violations. CHARMM27 produced structures with reduced exposure of the hydrophobic surface of the peptide to the solvent. The AMBER03 force field produced well conserved helical regions, and did not lead to many interproton distance violations. However, the resultant loss of inherent flexibility of the chain limited the exploration of the conformational space, preventing the sampling of the biologically
important structures in the course of the room temperature simulations. This limitation may be overcome by employing a method capable of crossing free energy barriers (Chapter 5). The OPLS-AA and GROMOS 53A6 force fields also produced unfavourable results, where the simulated structures had very high NOE energies, indicating a large number of distance restraint violations, mostly contributed by the loss of the helical structure.

It must be stressed, however, that even though some force fields considered here may not have reproduced all features of the normal behaviour of chain B within our simulated timeframe, it does not mean that this behaviour will not be observed if longer or more efficient simulations are conducted. Overall, we can conclude that the CHARMM27 and GROMOS 43A1 force fields appear to be the most suitable for studies of insulin behaviour using the “spontaneous” evolution techniques, such as the room temperature MD simulations presented in this chapter.
Chapter 5

5. Exploring the folding free energy landscape of insulin

5.1 Overview

The need for more sophisticated methods capable of exploring larger conformational space is becoming more evident with the inability of classical molecular dynamics to efficiently explore the conformational space and obtain free energies for complex processes such as protein folding. A recently developed technique called Bias-Exchange Metadynamics (BE-META) has shown great potential for sampling vast conformational space at reasonable computational expense.

In this chapter, the BE-META technique was applied to investigate the folding mechanism of insulin, one of the most studied and biologically important proteins. The BE-META simulations were performed starting from an extended conformation of chain B of insulin, using only eight replicas and seven reaction coordinates. The folded state, together with the intermediate states along the folding pathway was identified and their free energy was determined. Three main basins were found, separated from one another by a large free energy barrier. The characteristic native fold of chain B was observed in one basin, while the other two most populated basins contained “molten-globule” conformations stabilised by electrostatic and hydrophobic interactions, respectively. The implications and relevance of these findings to the folding mechanisms of insulin were investigated in detail and presented in this chapter.
5.2 Introduction

One of the most fundamental phenomena in nature is the capability of proteins to fold \textit{de novo} to their native conformation, also known as their biologically functional state. In the last couple of decades, significant advances have been made in the understanding of protein folding through experimental and theoretical approaches. However, to this day the folding process of a linear polypeptide strand to its three-dimensional, biologically active conformation is poorly understood. Insufficient sampling of the conformational space available to a biological system remains a problem for theoreticians even with the significant improvements in computer technology. The complexity and ruggedness of the free energy surface, comprised of numerous minima, induces difficulties in using classical MD for studying complex processes such as protein folding, as the system can easily get trapped in one of the local minima and fail to properly sample the rest of the conformational space. In order to overcome this complexity it is necessary to employ a methodology that is capable of accelerating rare events, specifically, configurational changes that involve the crossing of large free energy barriers. Few novel techniques capable of exploring wider conformational space have recently been developed, and one such technique is Bias-Exchange Metadynamics (BE-META). Using this technique, Piana and Laio were able to simulate reversibly the fold of Trp-cage protein with results revealing structurally related thermodynamic and kinetic properties, which were in agreement with the available experimental data [63]. In a recent study the effect of a point mutation, Pro62Ala, on the fold of villin and advillin was examined using BE-META, and predictions made from this work were validated with NMR and CD experiments [64].

In this chapter, BE-META and the accompanying analysis methods were utilised to investigate the folding mechanism of insulin, in particular the folding pathways of its chain B in explicit solvent. The structural characterisation and biological relevance of the conformations revealed by the simulations is described, accompanied by a fully atomistic model of the structural transitions and possible folding pathways of insulin chain B.

A description of the BE-META method is presented in Chapter 2, while the specific details regarding the collective variables used and the parameters applied for this study are stated in the \textit{Computational Details} section of this chapter. The model protein insulin was introduced in Chapter 3, with concise description of its structure and function, however
information and recent discoveries relevant to the folding of insulin are presented here also.

Insulin is composed of two chains, A (21 residues) and B (30 residues), linked together by two disulfide bonds. Chain B of insulin is believed to retain much of its structure independently of chain A [232-234]. A recent study by Budi et al. confirmed this property as they looked at the effect of thermal and chemical stress on isolated insulin chain B and complete insulin monomer [225, 229]. Structure-activity studies of insulin indicate the C-terminus of chain B as integral to receptor binding [248, 250, 310], and also suggest that the conformation of the C-terminus is influenced by the structure of the N-terminus [235, 238, 251]. The inherent flexibility of the N- and C-terminal regions of chain B of monomeric insulin was observed both by experimental and theoretical studies. Various investigations of insulin [236, 252], including that on a preliminary crystallographic structure of the native insulin monomer at low pH [253], and also on a solution structure of isolated chain B determined by NMR spectroscopy [232, 233], confirmed both termini’s mobility. Investigations performed by Legge and co-workers yielded information about the structure and dynamics of insulin with respect to its physiological behaviour by performing multiple MD simulations using the CHARMM27 force field in explicit solvent and ambient conditions [230]. Their work highlighted the importance of packing interactions for the conformational behaviour of chain B of insulin, specifically structures stabilised by localised hydrophobic interactions. Although insulin has been the subject of a large number of experimental and computational investigations (see Chapter 3; Insulin), studies of the monomeric or isolated chain B structure and dynamics in solution have been limited by the monomers’ susceptibility to self-associate into oligomers. To the best of our knowledge there is no other published study that has successfully simulated the folding of isolated chain B of insulin in explicit solvent.

5.3 Computational Details

For this study the Gromacs suite of programs was utilised [209] modified to perform bias-exchange metadynamics. The AMBER03 force field [165] was used for all calculations. We have previously shown (Chapter 4, [311]) that this force field “favours” helical structures and has a “rugged” energy surface. The AMBER03 force field was applied in this study to test the
ability of BE-META to explore the conformational space of insulin by overcoming these energy barriers.

The time step for all the simulations was set to 2 fs. Atom based cutoff of 8 Å was used for nonbonded van der Waals interactions. The Particle Mesh Ewald (PME) summation method [294] was applied to correct long-range electrostatic interactions. All bond lengths were constrained to their equilibrium value with the LINCS [193] algorithm. Constant temperature was achieved by coupling the system to a Nosé-Hoover thermostat [312] with a coupling time of 1 ps. Constant pressure was achieved by coupling the system to a Berendsen barostat [188] with a relaxation time of 4 ps.

The starting structure for this study was an arbitrary unfolded conformation sampled from our previous work on the effect of electric field [227] on the conformation of porcine insulin (PDB code 1ZNI [245]), see Figure 5.1. The protein was enclosed in a periodic box of 46 Å × 70 Å × 46 Å size, then solvated with 4220 TIP3P [313] water molecules, corresponding to water density of ~1.0 g/cm$^3$. The positive charge of the protein was neutralised by adding two Cl- counterions. The whole system was energy minimised to remove steric clashes using the steepest descent algorithm after which 200 ps of NPT molecular dynamics at 298 K and 1 atm were performed to equilibrate the protein and solvent.

![Figure 5.1](image_url) The starting structure selected for this study on the folding mechanisms of insulin chain B.

For the BE-META simulations seven generalised reaction coordinates were applied, none of which require a priori knowledge of the folded state. A neutral walker was also implemented, which is not biased by any metadynamics potential, i.e. evolves as a classical MD simulation, but is allowed to exchange conformations with the other replicas. The neutral walker statistics are approximately canonical as was shown by Piana and Laio in reference [63]. As the other trajectories in the BE-META calculations were generated under the
influence of history-dependent potential, a more complex analytical approach is required to exploit the resultant statistics. A methodology has recently been developed by Marinelli et al. to extract detailed kinetic and thermodynamic information of a complex system simulated under the influence of a time-dependent potential [65]. The algorithm behind this program is briefly summarised in the following sections of this chapter.

5.3.1 Collective variables

The seven uniquely defined collective variables (CV) implemented were: (1) the number of backbone H-bonds ($N_{hb}$), (2) number of Cγ contacts ($N_{C\gamma}$), (3) number of salt bridges ($N_{sb}$), (4) α dihedral fraction ($\Phi_{\alpha_1}$) applied to the 1st half ($^2$VNOHLCGSHLVEAL$^{15}$) of the protein, (5) α dihedral fraction ($\Phi_{\alpha_2}$) applied to the 2nd half ($^{16}$YLVCGERGFFYTPK$^{29}$) of the protein, (6) dihedral correlation ($\Phi_{corr_1}$) applied to the 1st half and (7) dihedral correlation ($\Phi_{corr_2}$) applied to the 2nd half of the protein. The terminal residues were excluded from these biases to enable their natural flexibility. The CVs (1-3) are defined as follows:

$$N = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(1 - \frac{r_{ij}^8}{r_0^8} - \frac{r_{ij}^{10}}{r_0^{10}}\right)$$

where $N$ is the number of contacts and $r_{ij}$ is the distance between atoms $i$ and $j$. The number of hydrogen bonds in $N_{hb}$ were calculated based on contacts between the HN and O backbone atoms of the protein. The contacts made between the Cγ atoms of chain B were counted by the $N_{C\gamma}$ variable. The potential salt bridge pairs, $N_{sb}$ were considered between the C of the carboxylic groups and the Nζ of Lysine and the Cζ of Arginine. The distance constraint $r_0$, for $N_{C\gamma}$, $N_{sb}$ and $N_{hb}$ was set to 5.0, 4.5 and 2.0 Å, respectively.

The CV4 and CV5 were defined as:

$$\Phi_{\alpha} = \sum_{i=1}^{N} \frac{1}{2} \left[1 + \cos(\varphi_i - \varphi_0)\right]$$

where $\Phi_{\alpha}$ is the helical content of the backbone, $\varphi_i$ is the backbone dihedral angle of residue $i$ and $\varphi_0 = -50^\circ$. 

102
While CV6 and CV7 were defined as:

$$\Phi_{corr} = \sum_{i=2}^{N_i} \sqrt{[1 + \cos^2(\phi_i - \phi_{i-1})]}$$  \hspace{1cm} (5.3)

where $\Phi_{corr}$ is the dihedral correlation, $\phi_i$ are the backbone dihedral angles of residue $i$.

These CVs are related to the free energy barriers between peptide conformations, as follows. The free energy barrier associated with the formation or disruption of H-bonds is described by the $N_{hb}$ variable. The backbone conformational changes are affiliated with the $\Phi_{a1}$, $\Phi_{a2}$, $\Phi_{corr1}$ and $\Phi_{corr2}$ variables. The $N_{C_\gamma}$ variable describes barriers associated with the formation of hydrophobic clusters, and the $N_{sb}$ variable describes barriers associated with the formation of salt bridges.

Each simulation was performed with 8 replicas, one neutral and one for each collective variable. Exchanges between replicas were allowed every 20 ps of MD simulation. Gaussian potentials of height 0.1 kJ/mol were added to the time-dependent potential every 500 steps (1 ps) during the entire MD simulation. The width of the gaussians (which ultimately determines the resolution of the free energy reconstruction) for each collective variable $N_{hb}$, $N_{C_\gamma}$, $N_{sb}$, $\Phi_{a1}$, $\Phi_{a2}$, $\Phi_{corr1}$ and $\Phi_{corr2}$ was chosen to be 2.0, 2.0, 0.5, 0.8, 0.8, 0.5 and 0.5, respectively. The rate and accuracy of exploration of the free energy surface is dependent on the chosen width and height of the gaussians, in the same manner as in the ordinary metadynamics [314]. The selected values were determined to give satisfactory results based on an extensive testing performed by the developers of the method [63]. Each replica was evolved for 96 ns, producing an accumulated total of 768 ns.

### 5.3.2 Analytical method for biased statistics: Kinetic model construction

An overview of the algorithm behind the construction of the kinetic model for our system is presented, however for in-depth information the reader is referred to reference [65]. The kinetic model implemented here has the form of a generalised rate equation, in which the nodes are a set of structures representative of all the states explored by the system. These structures are determined by the cluster analysis performed on the whole collective variable space. This analysis involves partitioning the seven dimensional reaction coordinate space into small hypercubes forming a regular grid. The structures within a single hypercube form a
cluster and the conformation near the hypercube centre represents the reference structure (RS) for this cluster. The size of the hypercube is defined by the side in each direction \( ds = (ds_1, ds_2, \ldots, ds_7) \). A convenient choice is \( ds_j = 2\sigma_j \), where \( \sigma_j \) is the width of the gaussian used for the CV \( j \). To check the cluster size consistency, cubic sides of \( 1.5\sigma_j \) and \( 2.5\sigma_j \) have been also attempted. The equilibrium probability to observe each structure is determined from the free energies obtained by the BE-META simulations. As these free energies are reconstructed as a function of one variable, while the nodes live in a multi dimensional space, the canonical weight of each cluster was estimated by a weighted histogram procedure (WHAM) \([315]\) based on the metadynamics bias potentials. Kinetic properties of the system are investigated by constructing a rate equation \([65, 316, 317]\) describing transitions between each pair of neighbouring nodes:

\[
k_{\alpha\beta} = k_{\alpha\beta}^0 e^{-\frac{1}{2kT} \left( F_\beta - F_\alpha \right)}
\]  

(5.4)

where \( k_{\alpha\beta}^0 = k_{\beta\alpha}^0 \) is a function of the cluster’s position and the diffusion matrix \( D \). The equation describes the diffusion between clusters with a barrier determined by their free energy difference. The diffusion matrix has been calculated for every cluster size using a maximum likelihood approach \([65]\). A 50 ns equilibrium MD trajectory, starting from the folded state has been employed for this purpose. The diffusion matrix calculated with the time lag of 0.5 ns has been chosen for kinetic analysis, as after this time variations in the matrix are negligible. The network model constructed in this manner is studied by the Markov cluster (MCL) analysis method \([318, 319]\). This technique enables the finding of metastable sets (basins) of the system, which gives a quantitative picture of the structure and energies of the conformational transitions of the peptide.

### 5.4 Results and Discussions

With the computational setup described above, BE-META simulation of chain B of insulin at 298 K were performed, starting from an extended conformation using 8 replicas. The compilation of structures accumulated from the neutral (non-biased) replica is initially discussed, followed by the investigation of the statistics collected from the biased trajectories.
CHAPTER 5. Exploring the folding free energy landscape of insulin

5.4.1 Neutral replica analysis

The statistics accumulated by the neutral replica are approximately canonical. Cluster analysis was performed on the neutral replica population using a simple backbone RMSD clustering method as described by Daura et al. [296]. A RMSD cutoff of 0.2 nm was applied on the backbone atoms of all pairs of structures used in the clustering method. After an initial transient period, the free energy profiles are approaching convergence and the population of the neutral replica starts stabilising. The cluster analysis of the last 30 ns of the neutral replica trajectory (Figure 5.2) allows an estimation of the free energy of the structures involved in the folding process. The total number of clusters explored by the neutral replica was 268. The free energy of each cluster \( E_i \) was determined using:

\[
E_i = -k_B T \log(N_i)
\]  

(5.5)

where \( k_B \) is Boltzmann’s constant, \( T \) is temperature and \( N_i \) is the total number of structures contained in cluster \( i \). The free energies of the 20 most populated clusters as well as their corresponding errors are presented in Figure 5.2. The errors were calculated as the standard deviation between block averages of 10 ns taken from the trajectory.

Figure 5.2 The free energy of the 20 most populated clusters. The native-like cluster is shown as an inset of the figure. The structures showing the most typical features are represented on the right of the graph. The clusters are numbered in the order of free energy with ascending cluster ID.
Several cluster reference structures were selected for their typical structural features and are shown on the right of Figure 5.2. The two most populated clusters 1 and 2 (~19% population) are molten globules, exhibiting native-like secondary structures with reduced level of amino acid packing as seen in the native state of chain-B. However, it can be noted that the inherent propensity of proteins to minimise the solvent exposure of non-polar residues has moved the termini close to one another. Insulin structures in a molten-globule-like state have been previously reported in the presence of 20% acetic acid [320]. Based on these findings, it was postulated that the lack of tertiary structural detail is intrinsic to the native insulin monomer and that the phenomenon per se is important for the interactions of insulin with its receptor.

Structures observed in cluster 11 (inset of Figure 5.2) have almost identical architecture with the experimentally observed Rf-state [245] of chain B: they contain a well conserved \( \alpha \)-helix, a \( \beta \)-turn and two extended termini. The RMSD of the backbone over the helical region (residues Ser9 to Cys19) between the reference structure of cluster 11 and the crystal structure of the Rf-state is 1.6 Å. A progressive helix formation in the region between residues Ser9 to Cys19 can be observed between clusters 2 to 15. This illustrates the generic protein behaviour, i.e. although proteins fold to minimise the exposure of nonpolar surface to water, it is the hydrogen-bonding interactions between buried backbone groups that stabilise the formation of the secondary structure elements such as \( \alpha \)-helix, \( \beta \)-sheets and tight turns [309].

The free energy surfaces of our seven collective variables for the last 30 ns of the simulations are depicted in Figure 5.3. The position of the lowest energy state within the neutral replica, defined as the average value of the collective variable calculated on the structures belonging to the most populated cluster is represented in blue. It is evident from the free energy profiles that the most populated state is either characterised by the free energy minimum or is very close to it, except in the salt bridges replica, caused by insufficient sampling due to an artefact in the simulation program at the border of a limited variable [321]. The statistics accumulated by the biased trajectories were further explored and a possible kinetic model of the folding mechanisms of chain B was proposed below.
Figure 5.3 Free energy profiles calculated from the BE-META simulations of chain B of insulin. The position of the lowest energy state of the neutral replica is marked by a blue square on each free energy curve. A typical lowest energy structure is also shown.
5.4.2 Kinetic model construction

A recently developed analytical approach [65], briefly summarised in the computational details of this chapter, was implemented to extract kinetic and thermodynamic information from our BE-META simulations on chain B of insulin. Cluster analysis was performed on the statistics accumulated in the last 40 ns of the simulation and ~17000 clusters were found using a cubic side $d_s = 2\sigma_i$. The population of each cluster was assigned by the WHAM procedure described by Marinelli et al. [65] and their free energies were determined (F, kcal/mol). Several values of filling time were examined, such as 30 ns, 35 ns and 40 ns. The results from these filling times were in good agreement with each other, yielding similar population for all of the most populated clusters with energy variation below $kT$. The filling time of 35 ns was used for the rest of the analysis, as after this time the most populated clusters appeared in the neutral replica, demonstrating a good compromise between the statistics achieved from the accurate sampling and the convergence required.

The kinetic model as implemented in reference [65] was applied to construct the transition matrix between clusters used in the MCL method, where diffusive dynamics is assumed between neighbouring clusters. The MCL method enables the determination of the metastable sets (basins) of the system. Clusters within each basin are structurally similar, and therefore the transitions between clusters that belong to the same basin are faster than transitions between different basins. The separation between basins is kinetic in nature, so there is an energy barrier separating one basin from another. The parameter $p$, used in the MCL method, controls the height of the energy barriers for clustering, so that a lower $p$ results in a smaller number of basins. The MCL calculation was performed with different values of the parameter $p = 1.08, 1.12$ and 1.14. Each $p$ was also applied to the free energies derived at the different filling times. We found that the population of the basins obtained from the various filling times is consistent and all most populated basins are maintained.

With $p = 1.08$ three basins were obtained with population of 70%, 25.5% and 4.5%, respectively. As they are obtained with a low $p$ value a high free energy barrier is expected between them. The most populated basin (molten-globule 1, Figure 5.4) contains clusters that show several salt bridges between Glu13, Glu21, Lys29, Arg22 and the terminal residues. The second most populated basin (molten-globule 2, Figure 5.5) is stabilised primarily by hydrophobic interactions and the last basin contains mostly structures of the native-fold nature. For $p = 1.12$, molten-globules 1 and 2 split into several sub-basins for a total of 11.
basins with population above 1%. At $p = 1.14$ the most populated basins were not fully divided so the results at $p = 1.08$ and $p = 1.12$ were chosen for structural characterisation and analysis.

The low free energy clusters for each of the three basins obtained for $p = 1.08$ are shown in Figure 5.4, 5.5 and 5.6, respectively. The most populated basin contains structures mainly governed by electrostatic interactions, where the charged side-chains of the most stable cluster (see Figure 5.4) are closely interacting with the terminal regions of the protein. In the low free energy clusters of molten-globule 2 (Figure 5.5a,b), residues Phe24 and Phe25 are packed into a compact hydrophobic core (less than 1 kcal/mol from the lowest free energy cluster). This observation implies that these two residues play an important role in the stabilisation of the basin. Furthermore, the Cys residues in these clusters are in favourable orientation for the disulfide bond formation with insulin chain A. In the less stable cluster Tyr16 can be found shielded from the solvent (see Figure 5.5c). Part of the central $\alpha$-helix is still retained in the structures within the molten-globule 2 basin. Possible roles of the above mentioned residues in the folding mechanism of chain B are discussed later. The most stable cluster of the native-like basin resembles the Rf-state of chain B (see Figure 3.2, Figure 5.6a, Figure 5.9). The N-terminal $\alpha$-helix can be found unfolded in some less stable clusters (see Figure 5.6b) belonging to the same basin or near the border of the native-like and molten-globule 2 basins. This structure resembles states T and O of chain B of insulin shown in Figure 3.2. Detailed structural comparison with experimental data was performed on the conformations contained in the folded state basin and the results are presented below.
CHAPTER 5. Exploring the folding free energy landscape of insulin

Figure 5.4 Representative low free energy cluster of the three basins determined using $p = 1.08$. **Molten-globule 1** Most stable cluster of the basin, where all charged residues interact with the N- and C-termini of the protein, giving rise to strong electrostatic contacts.

Figure 5.5 **Molten-globule 2** a) Most stable cluster of the basin, where Phe25 (green) is buried in a hydrophobic pocket, while Phe24 (yellow) is partially exposed to the solvent. b) This structure is only 0.9 kcal/mol from the most stable cluster and shows the presence of a hydrophobic core in which Phe24 (yellow) is present. Phe25 (green) is partially solvent exposed in this basin. c) In this cluster the Tyr16 (orange) residue and partially Phe24 (yellow) are buried in a hydrophobic pocket. The hydrophilic part of Tyr16 is directed toward the Arg11, Gln21, and the terminal salt bridges.

Figure 5.6 **Native-like** a) Lowest free energy cluster in the folded state basin. b) Cluster present between the native-like and molten-globule 2 basin having the N-terminal $\alpha$-helix unfolded.
Thermodynamic properties of the three basins are listed in Table 5.1. All observables and error evaluations are calculated according to the method described in reference [64, 65]. Both molten-globules 1 and 2 show an enthalpic penalty with respect to the native-like basin, although for molten-globule 1 the difference is only slightly larger than the standard deviation. The two molten-globules are entropically stable. As expected the basin with the largest entropy is molten-globule 2, since it has the highest content of hydrophobic contacts. In Figure 5.7 the occupancy of each basin as a function of temperature is reported. According to its entropy content, the population of molten-globule 2 increases as temperature is raised. The native-like basin occupancy does not vary strongly with the temperature and the two molten-globules show complementary behaviour. The presence of this type of conformation is supported by NMR studies performed by Hua et al. which found local differences between the solution structures of insulin and the crystal structure, suggesting that the biologically active form of insulin is a molten-globule [320, 322].

<table>
<thead>
<tr>
<th></th>
<th>Molten-globule 1</th>
<th>Molten-globule 2</th>
<th>Native-like</th>
</tr>
</thead>
<tbody>
<tr>
<td>% occupancy</td>
<td>70.0 ± 8.0</td>
<td>25.5 ± 7.0</td>
<td>4.5 ± 1.0</td>
</tr>
<tr>
<td>ΔH (kcal/mol)</td>
<td>1.4 ± 0.3</td>
<td>7.0 ± 1.0</td>
<td>0.0 ± 0.8</td>
</tr>
<tr>
<td>ΔS (kcal/mol K)</td>
<td>0.010 ± 0.0015</td>
<td>0.027 ± 0.004</td>
<td>0.0 ± 0.003</td>
</tr>
</tbody>
</table>

*Table 5.1* Thermodynamic properties of the basins found using the MCL algorithm with \( p = 1.08 \). All thermodynamic properties were calculated with respect to the native-like basins.
CHAPTER 5. Exploring the folding free energy landscape of insulin

Figure 5.7 Occupancy (%) of the three basins found for $p = 1.08$ as a function of temperature. A simple linear extrapolation [65] from the calculated cluster enthalpy and entropy at 300 K has been used to obtain the basin occupancy versus temperature.

Figure 5.8 shows the eight most populated basins obtained for $p = 1.12$, represented in three dimensions, as contours around the most populated state: the $\gamma$ contacts, the $\alpha$-dihedral fraction in the 1st half of the protein (2VNQHLCGSHLVEAL$^{15}$) and the $\alpha$-dihedral fraction in the 2nd half of the protein (16YLVCGERGFFYTPK$^{29}$). Each colour corresponds to a different basin. The lowest free energy cluster of each basin is represented as a sphere of the same colour and its 3D structure is depicted in order of increasing energy. Transitions among these basins are expected to be faster compared to the ones at $p = 1.08$. 
Figure 5.8 Metastable sets (cluster basins) detected by MCL using $p = 1.12$, are shown as coloured contours, with the coloured sphere corresponding to the lowest free energy cluster of each basin. The respective structures are presented above with the same colour code and ordered based on their free energy (1 has the lowest, 8 has the highest free energy).
Although the most populated cluster in our simulations is not the native state of chain B, we did sample structures resembling the R and Rf-states presented in Figure 3.2. It is evident from Figure 5.8 that there is a large energy barrier i.e. a kinetic gap, between the folded state basin (basin 5) and the other partially folded basins. The structural stability of the folded state could depend on factors such as the binding to chain A. Specifically, insulin’s three disulfide bridges (A6-A11, A7-B7, and A20-B19) play a critical role in the protein synthesis, structure and stability. To investigate the effect of non-native disulfide pairing on insulin’s structure and biological activity, Hua et al. prepared by direct chemical synthesis two insulin isomers having disulfide bonds between the following pairs: (1) A7-A11, A6-A7, A20-B19 and (2) A6-A7, A11-B7, A20-B19 [323]. Using CD and NMR spectroscopy they found that the engineered isomers have less helical content compared to native insulin. Their thermodynamic studies by CD-detected guanidine denaturation demonstrated that their non-native disulfide paired isomers are markedly less stable than the native insulin, suggesting that this instability is in qualitative agreement with the isomers’ lower α-helix content.

5.4.3 Structural analysis of the folded state basin

Structural analysis was performed on the folded structures identified in basin 5 of Figure 5.8 which exhibit conformational elements typical for the X-ray crystallographic states of chain B shown in Figure 3.2. The equilibrium averages of the observables considered were calculated using the free energies obtained by the WHAM procedure. The average value of the observable $O$ is given by

$$
\left\langle O \right\rangle = \frac{\sum_n O_n e^{-F_n/T}}{\sum_n e^{-F_n/T}}
$$

(5.5)

where the sum is performed over all clusters, $T$ is the temperature, $F_n$ is the free energy and $O_n$ is the average value $O$ in cluster $n$.

To the best of our knowledge the structure of isolated chain B of porcine insulin has not been determined. Most of the published work has been performed on engineered insulin monomer [324] and NMR studies have been presented on mutated [233] and oxidised [232] isolated chain B. A comparison of the structures contained in our simulated folded state basin has been performed with the solution NMR structure of isolated chain B of insulin [232],
X-ray crystallographic structures [245] and previous molecular dynamics studies on chain B which reproduced experimental conformations [230, 311]. Conserved structural features of chain B were identified and their RMSD was calculated from the crystallographic Rf-state. A potential for disulfide bridge formation was investigated by calculating the solvent accessible surface (SAS) area of the Cys residues. The interproton distance violations of the folded structures based on nuclear Overhauser enhancement (NOE) distance restraint data were also analysed in detail.

The data derived from the NMR structure of bovine insulin chain B in solution provides a useful benchmark for evaluating the conformations sampled during the BE-META simulations. Hawkins et al. derived NOE constraints from a 250 ms NOESY spectrum of the oxidised chain B at 500 MHz, 300 K and pH of 2.2 to 2.5 [232]. As coordinates for the structure were not available, NOE distance restraints were used for validation of the conformations obtained. In the folded state basin, on average 13 out of 309 restraints were violated, with only 6 severely over their upper limit (violation > 1.0 Å). A mean interproton distance violation of 0.13 Å was also calculated. Most violations were found at the N- and C-terminal regions. Interestingly, these results are almost identical to the results obtained in our recent investigation, where two independent MD simulations were performed using the AMBER03 force field, starting from the folded crystal structure of chain B (Chapter 4 and [311]). The same NOE distance restraints, as implemented in this study, were used to validate the structures sampled. The total number of violations in the two individual simulations was 13 and 9 respectively, with 7 and 6 violations considered severe. Several studies have alluded to the difficulty in the structural definition of the C-terminal region of chain B due to its inherent flexibility [324]. Moreover, studies have shown that receptor binding must be accompanied by a major conformation change in the carboxyl terminus of chain B [308].

To further inspect the conformations from the folded state basin, the average equilibrium backbone RMSD was calculated of the important structural elements of chain B, such as the α-helical region (residue Ser9 to Cys19) and β-turn region (residue Gly20 to Gly23). The central helix of chain B plays a key role in insulin’s activity, while the conservation of the β-turn between residues Gly20, Glu21, Arg22 and Gly23, plays an important part in the folding and conformation stability [325]. The Gly20 to Gly23 turn is integral to the chain B secondary structure because it enables the C-terminal β-strand to pack against the central α-helix. Although the turn is more flexible than these adjoining structural elements, its pattern of hydrogen bonds and dihedral angles is essentially identical among
multiple crystal forms [235, 237, 238, 240, 326]. Therefore, RMSD calculations were performed from the crystallised Rf-state of chain B to investigate the stability of the helix and β-turn within the folded state basin. An overlay of the structure of the lowest energy cluster from the folded state basin (structure 5, Figure 5.8) and the crystal structure of chain B, taken from the PDB code 1ZNI is depicted in Figure 5.9, aligned over the central helical region.

Figure 5.9 The lowest energy structure (blue) from the folded state basin superimposed with the X-ray crystallographic structure of chain B, PDB code 1ZNI (pink).

The average RMSD over the helical region in the folded state basin is 2.2 Å, where that for the lowest energy state is 1.7 Å. This result is illustrated by the observed alignment between the lowest energy state and the crystal structure, shown in Figure 5.9. Excellent agreement was also found with the RMSD calculated from our classical MD calculations performed on the folded crystal structure, where an average RMSD of ~1.9 Å was obtained for 2 independent 50 ns simulations (Chapter 4 and [311]). The turn region between residues Gly20 and Gly23 was calculated to have an average RMSD of 1.8 Å, with the lowest energy cluster having deviation of 1.3 Å. Overall, these results are in agreement with experimental and theoretical data obtained by a different approach, as well as with the structural analysis performed in this work on the neutral replica statistics.

To examine the relative exposure to water of residues Cys7 and Cys19, which bind chain A with B via two disulfide bonds, their SASA was calculated from the unfolded (structure 1 and 2, Figure 5.8) and folded (structure 5, Figure 5.8) basins. For the folded basin we found that Cys7 had low SASA of 4.6 Å², while Cys19 was more solvent exposed with SASA of 10.4 Å². Interestingly, this result is in a very good agreement with the reference
crystallographic Rf-state (1ZNI.pdb) of chain B having SASA of Cys7 = 4.3 Å² and Cys19 = 12.2 Å². In the most populated “molten-globule” states 1 and 2, the cysteine residues were similarly exposed to the solvent, having the SASA of 9.9 Å², 8.9 Å² for Cys7, and 8.8 Å², 7.9 Å² for Cys19, respectively. This result suggests that although chain B is isolated, the disulfide forming residues are in a favourable arrangement for bonding to chain A.

The key elements in directing insulin-receptor interactions and in formation of insulin dimers are the well conserved phenylalanine residues at B24 and B25. Mutational analysis has been extensively applied to investigate the importance of these residues in the hormone insulin (see references in Shoelson et al. [327]). The conservation of secondary structures and support in intermolecular association are two distinct roles associated with the benzyl side chains of Phe24 and Phe25 [219, 297, 298, 328]. It has been suggested that Phe24 interacts with the hydrophobic core of insulin, specifically with B12, B15 and the A20-B19 disulfide bond. Furthermore, Phe24 side-chains have been shown to stabilise the β-turn at B20-B23 [308], however Phe24 may not interact directly with the receptor, unlike Phe25 which is more exposed to the solvent and is easily accessible for receptor interaction. We support these findings by the observed conformations of Phe24 and Phe25 in the folded state (structure 5, Figure 5.8). A turn is formed at the location of the two aromatics as a result of a salt bridge formed between Ala30 and Arg22, resulting in Phe25 being slightly solvent exposed. Aromatic ring interaction between Phe25 and His5 is preserving the partial packing of the C-terminal β-strand against the α-helix. Furthermore, residue Phe24 is strongly interacting with the α-helix and reduces the flexibility of the β-turn. These observations are in agreement with the proposed roles of Phe24 and Phe25, and give insight into the possible structural transformation the C-terminus adopts upon folding.

5.4.4 Folding pathway of chain B of insulin

The dynamics of the system was investigated by applying the Kinetic Monte Carlo (KMC) [329] method to compute the transition between the basins found by the Markov cluster analysis. The method as described in reference [65] was applied. This enabled a construction of a reduced rate model in which only transitions between the basins are considered. The rate constants for this model are given by the inverse of the transition times. For example, by taking two basins A and B with occupancy, $P_A$ and $P_B$, the rate constant to go from A to B is calculated considering the number of times ($N_{AB}$) a trajectory goes from A to B without
passing any other basins during a long KMC simulation ($\tau_{\text{KMC}} = 1.5$). In this way the actual time of transition to go from A to B is estimated as $k_{AB} = N_{AB} / (P_A \times \tau_{\text{KMC}})$. To minimise the number of recrossing only the stable clusters of each basins were considered, i.e. ~70% of the basin population. The kinetic scheme of the insulin chain B folding is outlined in Figure 5.10.

![Figure 5.10 Schematic representation of the basin dynamics. Transitions between the basins are represented by an arrow along with the corresponding transition time. Transitions that occur on a time scale longer than 1 µs are represented as dashed arrows.](image)

The basins are organised in a similar arrangement to their contour representation in Figure 5.8. The results show that the folded basin is connected directly only to the molten-globule 2 basin, while molten-globule 2 basin is connected with that of molten-globule 1, thus forming an overall linear pattern (see Figure 5.10). Transitions between these basins occur in a few thousands of nanoseconds, suggesting that the residence time of the three basins is of the order of several microseconds. Transitions between sub-basins (obtained with $p = 1.12$) inside each well (obtained with $p = 1.08$) occur in tens or hundreds of nanoseconds. The transition region between the folded state and the molten-globule 2 is made of states that contain unfolded N-terminal helical region, similar to the T- and O-state of chain B (Figure
3.2, Figure 5.6b). This indicates that the first step of going from the folded state to the nearest molten-globule state is by unfolding the N-terminal α-helix (Figure 5.6a,b). In the second step the N-terminal region interacts with the remaining helical elements forming a compact intermediate in which the Tyr16 is buried in a hydrophobic pocket (Figure 5.4d). All these transitions involve the high free energy clusters (e.g. 3-5 kcal/mol from the most stable cluster). The last step involves a partial exposure of Tyr16 to allow Phe24 to be packed via hydrophobic contacts (see Figure 5.5b), which results in a gain in free energy. The transition from molten-globule 2 to 1 seems to involve the intermediate structure presented in Figure 5.5c. In fact, the state displayed shows a salt bridge formation at the chain’s terminal regions. The last step involves the expulsion of Tyr16 from the pocket forming the electrostatic core (Figure 5.6b). It is noteworthy that many of the clusters associated with the insulin chain B folding could not exist in the presence of disulfide pairing with chain A as this would result in a loss of flexibility. Nevertheless several stable clusters in the molten-globule 2 basin contain Cys residues that are solvent exposed and at the right orientation to form disulfide bridges. This suggests that this basin could exist even after binding with chain A in the form of a two state model for the protein [330].

5.5 Conclusions

As presented in Chapter 2 of this thesis, time evolution techniques such as molecular dynamics are limited in their ability to efficiently sample the conformational free energy surface of a protein, thus making them unsuitable for applications to complex processes such as protein folding. Fortunately, methodologies capable of crossing large free energy barriers at a limited computational cost are becoming increasingly available.

In this chapter explicit solvent bias exchange metadynamics simulations were performed to effectively sample the conformational space available to chain B of insulin and to shed light on the complex structural transitions this important protein undergoes upon folding. To exploit the statistics accumulated using this powerful technique, a recently developed analytical method was used to construct a model describing the complex conformational transitions chain B experiences. The model suggested an existence of three metastable basins separated by large free energy barriers. The two most populated basins had structures with molten-globule characteristics, one being governed by electrostatic
interactions and another primarily by hydrophobic contacts. This finding is supported by experimental studies which suggested this type of conformation to be biologically active [320, 322]. The third basin is comprised of conformations with folded structural elements, resembling the known crystallographic states of chain B (α-helix, β-turn and flexible termini). The folded state basin contained physiologically important features, such as: a well conserved α-helical region, a β-turn stabilised by interaction between Phe24 and the hydrophobic core of the α-helix; and structural transformation of the C-terminus, favourable for possible binding to the insulin receptor. All of the low energy structures from each basin had solvent exposed cysteine residues open for possible binding to chain A.

The MCL and rate calculations presented in this work show a three state model for the folding pathway of insulin chain B. Starting from an extended structure, at first the protein is governed by electrostatic interactions (molten-globule 1, Figure 5.4). A progressive building of hydrophobic core is initiated by the burial of the Tyr16, followed by further packing of Phe24 and Phe25 (molten-globule 2, Figure 5.5a,b), resulting in stable compact structures. Furthermore, the hydrogen bonding interactions between the buried backbone groups commence the formation of an α-helix at the core of the protein. An unfolded N-terminal region is found in the structures at the border of molten-globule 2 and the folded basin, suggesting that the last stage of the folding of chain B is the complete formation of the α-helix. The transformation from molten-globule 2 to a folded state requires crossing of a high energy barrier, and as a consequence of this tens of microseconds are required to make this transition. The calculated transition times gave further insight into the dynamics between the three basins, suggesting that the residence time of the three basins is of the order of several microseconds. We believe that the native disulfide pairing of chain A with chain B (A7-B7 and A2-B19) plays an important part in the stability of the α-helix, which effectively prevents the protein from unfolding and becoming a molten-globule.

The ability to obtain a reliable model of the folding pathway of a protein is governed by the method’s capability to sufficiently sample the conformational space, and of course the complementary analysis that will interpret the collected data. The need for further research and development of sophisticated techniques, such as BE-META is highlighted by the investigation presented, as complex processes such as enzyme reaction, protein association and aggregation will be accessible at reasonable computational cost.
Chapter 6

6. Investigation on the effects of mutation of the amyloidogenic apoC-II(60-70) peptide

6.1 Overview

Protein aggregation is increasingly becoming identified as the underlying cause of many debilitating diseases affecting large numbers of the world’s population today. A detailed investigation of the initial mechanisms of fibril formation and environmental influences is crucial in order to improve the understanding of this problem and to aid the development of therapeutics.

In this chapter the effects of mutation on the structure and dynamics of the amyloidogenic peptide apoC-II(60-70) were identified. Methionine (Met60) substitutions by hydrophilic Gln and hydrophobic Val residues were investigated. The results were compared with the recent findings from a study on the wild-type and methionine oxidation product of the apoC-II(60-70) peptide.

Structural evolution and cluster analysis were performed on all systems. Radius of gyration, solvation properties, side-chain interactions and the aromatic angle orientation were analysed in detail. Finally, the conformational features relevant to amyloidogenic propensities of the peptide were identified.
6.2 Introduction

An interesting phenomenon observed in protein behavior is the misfolding of the protein, either during the folding process or by unfolding from its native state under physiological conditions. In the misfolded conformation proteins can self-associate to form amorphous compounds or elongated structures, known as amyloid fibrils. The deposition of amyloid fibrils in the human body has been linked to many debilitating diseases, thus making this field of research important and urgent. Several models have been proposed for the mechanisms of fibrillation (discussed in Chapter 1), however the initial transformation of the protein monomer and subsequent self-association is poorly understood.

Conventional experimental techniques such as solution NMR and X-ray crystallography are limited in revealing structural information on the protein aggregates because they are usually large, insoluble and non-crystalline. The application of computational methods to study aggregation mechanisms enables the investigation of the initial stages of fibril formation at atomistic resolution. Specifically, the conformational changes experienced by the protein monomers under various conditions can be identified. This is important, as it will give insight in the initial process of protein structural transitions that may lead to subsequent association with other free monomers. Recent studies have suggested that the cytotoxic and disease forming species are the prefibrillar intermediates [126]. A possible therapeutic strategy against amyloid diseases could be the application of specific mutant variants that prevent the nucleation of the fibrils or assist in their dissociation. Detailed knowledge of the structure of the respective species at different stages of formation is essential in order to design such compounds.

Residue substitution methods have been applied in many studies of amyloidogenic proteins to determine specific amino-acid contributions to fibril formation, stability and dissociation. Specifically, the effects of oxidation have raised an interest due to its reversibility and promising inhibitory results. Maleknia et al. investigated the effects of amino acid side-chain oxidations by comparing the kinetics of fibril formation of oxidised and unoxidised proteins [331]. The wild-type (WT) and Val30Met Transthyretin (TTR) mutant were allowed to react with hydroxy radicals and other reactive oxygen species. They found that oxidation inhibited the initial rates and extent of fibril formation for both the WT (~76%) and Val30Met (90%) TTR proteins. The interference with fibril formation by oxidised Met
was also observed for proteins such as immunoglobulin light chain LEN [332], α-synuclein [333] and prion protein (PrP) [334].

Triguero et al. applied computational methods to investigate the effects of oxidised Met on the secondary structure of full-length Aβ(1-40) and Aβ(1-42) monomers [335]. They observed significant changes in secondary structure, from a stable β-hairpin found in the wild-type Aβ monomer to random coil elements in the oxi-Met35-Aβ(1-40). Reduction in the interaction between the N- and C- termini was also observed.

Hydrophobic effects can be considered the main driving force determining aggregation and fibrillation. Zagorski and co-workers suggested that in the early stages of Aβ-assembly, hydrophobic and bend-like structures are required to produce amyloid fibrils [336]. More importantly, they noticed that oxidation of Met35 blocks the critical hydrophobic association in the initial transformations. Oxidation of the Met35 side-chain to the sulfoxide derivative introduces a variation in the charge and electrostatic properties, resulting in a net increase in polarity of the hydrophobic C-terminus, eventually causing a reduced tendency to aggregate via a hydrophobic effect.

A recent study by Steinmetz et al. used a specially designed ccβ-Met peptide to provide an explanation as to why methionine oxidation can have such a significant effect on amyloid fibril formation [337]. Based on their findings they suggested that a certain number of site-specific hydrophobic interactions within the polypeptide chain are responsible for the formation of highly stable amyloid structure. Thus perturbation of one of these residues by the insertion of a single oxygen moiety can have a drastic effect on the kinetics and dynamics of fibril assembly.

It was recently discovered that human apoC-II protein readily aggregates into twisted ribbon-like fibrils at physiological pH and without prolonged agitation [277]. The effects of single residues and their mutation on the fibril forming propensities of native apoC-II and smaller apoC-II derived peptides have been investigated experimentally [281, 284, 285]. The full-length apoC-II protein contains two Met residues, one found at the beginning of the amino acid sequence (position 9) and the other in the fibril forming core (position 60). Recent experimental work explored the effect of oxidation and mutation of these residues on the amyloidogenic nature of the full-length protein [281]. The results demonstrated significant inhibition of aggregation when Met60 was oxidised, while a smaller effect was seen upon oxidation of Met9. Mutation studies of these residues showed that when Met60 was replaced
CHAPTER 6. Investigation on the effects of mutation of the amyloidogenic apoC-II(60-70) peptide

by Val the protein formed fibrils at a similar rate and morphology as the wild-type. Mutation of Met60 to Gln also leads to formation of fibrils, however much higher peptide concentration was required.

Short peptide fragments composed of apoC-II residues 60 to 70 have been suggested to be the driving fibril forming sequence in apoC-II [282]. A recent experimental and computational study performed by Hung and co-workers investigated the influences of phospholipids, methionine oxidation and acidic pH on the derived apoC-II(60-70) peptide [285]. They found that at fibril-favouring conditions (neutral and low pH) the peptide preferentially adopts structures similar in shape to a β-hairpin, while under fibril-disruptive conditions (lipid-rich and oxidised Met) significantly different structures were obtained. The formation of β-hairpin structures in relatively short amyloidogenic peptides has been observed in several peptides, such as LSFD [338], amyloid-β(25-35) [339], and in an 11-residue fragment of β2-microglobulin [340]. The development of β-hairpin structures in amyloidogenic peptides suggests this type of conformation to be a possible intermediate state a peptide monomer exhibits prior to association, thus making this structure favourable for peptide aggregation.

In a more recent study, experimental and computational techniques were utilised to explore the effect of oxidation and mutation of Met60 to Val and Gln on the stability of the apoC-II(56-76) peptide [284]. The wild-type apoC-II(56-76) and 56-76Met60Gln peptides readily assembled into fibrils with similar lag phase, whereas the apoC-II(56-76) with oxidised Met formed fibrils with much slower kinetics and was not completely inhibited. The slowest to form fibrils was the 56-76Met60Val peptide, which exhibited totally different aggregation kinetics compared to the full-length apoC-II with the same mutation. Legge and co-workers observed that the behaviour of the apoC-II(56-76) peptide system is different to that of the full length protein system, suggesting that the mechanism of fibril formation may also be different [284].

This work was extended by performing molecular dynamics simulation on the shorter amyloidogenic apoC-II(60-70) peptide. Specifically, the effects of various mutations on the conformation of the apoC-II(60-70) monomer were investigated. Single-point residue substitutions were performed at Met60 to hydrophobic Val (Met60Val) and hydrophilic Gln (Met60Gln) in order to compare with the behaviour of the wild-type and oxidised apoC-II(60-70) peptide. The dynamics and structures of individually mutated apoC-II(60-70) peptides
CHAPTER 6. Investigation on the effects of mutation of the amyloidogenic apoC-II(60-70) peptide

were compared with the conformations of wild-type and oxidised apoC-II(60-70) [285]. The Gln amino acid was chosen because of the similar hydrophilicity to oxidised Met, while substitution of Met to Val enabled the hydrophobic influences on the peptide’s structure to be examined. The conformational features and interactions of various mutations that may have inhibitory or amyloidogenic propensities were identified. Additionally, recent findings showing differences in the fibrillation kinetics between the full-length and derivative apoC-II(56-76) peptide were used for comparison with the structural studies of apoC-II(60-70). This allows the determination of the different mechanisms that are involved in fibril formation by apoC-II and its derivative peptides.

6.3 Computational Details

In this study the Gromacs 3.3 [209] simulation package was employed. The calculations were performed under NPT conditions with the united-atom Gromos force field parameter set 43A1 [174]. The oxidised Met side-chain force field parameters were derived from those of dimethyl sulfoxide (DMSO) for the Gromos96 force field. The time step for all the simulations was set to 2 fs. The electrostatic and Lennard-Jones interactions were calculated with a cutoff of 10 Å. The Particle Mesh Ewald (PME) [294] method was used to treat long-range electrostatic interactions. All bond lengths were constrained to their equilibrium value with the LINCS [193] algorithm.

The selected starting structure was the native conformation of the 11 residue sequence (MSTYTGIFTDQ) extracted from the NMR structure of apoC-II (PDB code 1SOH). Single point mutations were performed at Met60 to Met60Val and Met60Gln, and their structural representations are shown in Figure 6.1.

Each peptide was enclosed in a periodic box of 43 Å × 34 Å × 34 Å dimensions, and solvated with ~1600 SPC water molecules, corresponding to water density of ~1.0 g/cm³. The whole system was energy minimised to remove steric clashes using the steepest descent approach. To allow the solvent to relax around the protein 1 ns equilibration was performed at constant pressure (NPT) with the protein restrained. Simulation conditions of constant temperature at 300 K and constant pressure of 1 bar were achieved by coupling the system to
a Berendsen thermostat and barostat, respectively [188]. Following the equilibration stage, data collection for analysis was performed in an NPT ensemble for 650 ns for each mutation.

To allow a comparison of the effects of Met60Val and Met60Gln mutations with the wild-type and oxidised apoC-II(60-70) peptide, some results were reproduced from a recent publication from our group [285]. In particular, the secondary structure evolution plot and aromatic angle distribution graphs are shown in Figure 6.2 (a, b) and 6.6 (red, black), respectively, together with the results obtained in this study (Figure 6.2 (c, d), Figure 6.6, (blue, green). Additional analysis was performed on the wild-type and oxi-Met apoC-II(60-70) peptides, such as the radius of gyration over the equilibrium trajectory (Figure 6.4) and the most frequently sampled structures were identified using an RMSD based clustering method (Figure 6.3). Furthermore, supplementary calculations were performed on the structures contained within the most populated cluster of each system to allow direct comparison between the “equilibrium” conformations obtained of the wild-type and mutated peptides. Specifically, the solvent accessible surface area (SASA), root mean square deviation (RMSD), residue contact maps and termini distance separation was calculated for each peptide. The results from the abovementioned analysis are discussed in the Results and Discussions section of this chapter.

**Figure 6.1** Structure representation of apoC-II(60-70) monomeric peptide showing mutated residues and simulation times.
6.4 Results and Discussions

6.4.1 Structural evolution analysis

The secondary structure dynamics reveal details of the conformational changes experienced by the protein over the entire simulation period. Specifically for this study, the structural evolution analysis was very useful because the peptides under investigation differ by a single residue only, so it was possible to examine the effect of each mutation on the structure and dynamics of the peptide. The secondary structure was classified using the algorithm STRIDE which utilises hydrogen bond energy and mainchain dihedral angles in addition to bond distances [210]. The structural evolution plots of the wild-type and oxidised apoC-II(60-70) were reproduced with permission from the authors [285], and together with those for the mutated apoC-II(60-70) systems are presented in Figure 6.2.

The wild-type apoC-II(60-70) peptide exhibited stable conformations with well-defined structural features (Figure 6.2a). In particular, the β-hairpin motif was identified throughout the simulation. The β-hairpin structure was defined by two strands, Ser61-Tyr63 (strand 1) and Phe67-Asp69 (strand 2), that are adjacent in primary structure and form a loop in an anti-parallel arrangement (where the N-terminus of one strand is adjacent to the C-terminus of the other strand). The anti-parallel strands were linked by a short turn region composed of three residues, Thr64, Gly65 and Ile66. Similar turn structure was identified in an MD study of the apoC-II derived peptide made of residues 56 to 76 [341]. Legge et al. postulated that this turn region is an important feature formed during the nucleation stage of fibril formation due to the formation of a stable hydrophobic core. As discussed previously, recent experimental data has shown that wild-type apoC-II(60-70) peptides readily aggregate into fibrils [285], therefore the turn structure observed in the dynamics of this peptide potentially provides the conditions required for aggregation.

Inspection of the 630 ns simulation of the oxidised Met (oxy-Met) apoC-II(60-70) peptide showed several important changes in the secondary structure (Figure 6.2b). A stable core turn region formed between residues Tyr63 and Ile66 in the initial ~250 ns of simulation, followed by a shift of the turn to residues Thr64-Phe67. This structural element was supported by the formation of a strong hydrogen-bond bridge between these residues. In contrast, the three residues at each peptide terminus fluctuated between turn and coil conformations, indicating no significant interactions between these regions of the two strands. Overall, the
CHAPTER 6. Investigation on the effects of mutation of the amyloidogenic apoC-II(60-70) peptide

oxidised peptide exhibited higher structural flexibility compared to the wild-type apoC-II(60-70) peptide, with less hydrogen bonding and electrostatic interactions between strands 1 and 2. The different structural features of the wild-type and oxi-Met peptides observed computationally, may help explain their different aggregation behaviour observed in recent in vitro studies [285].

Distinct structural features were observed in the Met60Val mutated system compared to the oxidised Met, with the mutated structures showing greater similarity in conformation to the wild-type peptide. Throughout the simulation the Met60Val mutated peptide exhibited the β-hairpin conformation, where the core turn region between residues Thr64 and Ile66 was maintained by the electrostatic interactions between the two β-stranded chains, represented by yellow in Figure 6.2c. After ~320 ns of simulation the whole peptide displayed a turn conformation followed by a partial fold into a π-helix. This conformation lasted for ~17 ns, after which the peptide folded back into a β-hairpin that remained for the rest of the simulation. These observations further highlight the peptide’s preference for the β-hairpin conformation.

Increased variation in the secondary structure was seen in the Met60Gln system compared to the Met60Val mutation (Figure 6.2d). A stable β-hairpin structure was observed for the initial ~250 ns, with a core turn region between Thr64 and Gly65. This was followed by a shift of the core turn to residues Gly65-Ile66. Further increase in flexibility was observed by an extension of the central turn of the peptide to the region between residues Tyr63 and Phe67. This β-hairpin conformation was stabilised in the last ~100 ns of the simulation by hydrogen bonds between the two Thr residues at positions 62 and 68, respectively. A short-lived π-helix fold was noted when there was a reduction in interaction between the terminal residues of the peptide and an extension in the overall conformation was observed. This type of development was also identified in simulations using the Charmm27 force field on the same point mutations of the apoC-II peptide, suggesting the π-helix to be a consequence of these mutations rather than a force field artifact. Interestingly, no π-helix formation was observed in the oxidised Met simulation, possibly due to the well maintained central hydrophobic core restricting the extension of the peptide.

Even though some increased flexibility in secondary structure was observed in the Met60Val and Met60Gln mutated peptides compared to the wild-type, overall these systems maintained the β-hairpin motif throughout the simulations, mostly stabilised by hydrogen
bonding and electrostatic interactions. It can be suggested that the preservation of the hydrophobic core may provide the initial conditions for aggregation. However, the increased flexibility and reduced hydrogen bonding and electrostatic interactions between the terminal regions observed in the oxidised Met simulation, may limit the peptide from experiencing and maintaining the conformational transformations required for the formation of fibrils.

**Figure 6.2** Secondary structure evolution plot for each mutation of apoC-II(60-70) peptide; a) Wild-type; b) oxi-Met; c) Met60Val; d) Met60Gln. The secondary structure colour codes: magenta – α helix, red – π helix, cyan – turn, white – coil, yellow – extended conformation and green – hydrogen bridge.
6.4.2 Cluster analysis

To examine the most frequently occurring conformations in each simulation, an RMSD based clustering method was utilised. For the grouping of simulated structures in the cluster method, an RMSD cutoff of 0.15 nm was applied to the whole molecule. The total number of the identified clusters, along with the two most populated structures and their corresponding population are shown in Figure 6.3.

The cluster analysis of the wild-type apoC-II(60-70) peptide identified 1259 distinct clusters. Close inspection of the representative structure from the most populated cluster shows the presence of a $\beta$-hairpin conformation. The most populated cluster contains 50% of the structures generated by the simulations, highlighting the significant stability of this type of structure, compared to only 5.2% of the second most populated cluster.

The greater flexibility observed in the secondary structure analysis of the oxi-Met simulation, was reflected by the increased number clusters classified, 1417, where the most populated cluster contained 19.4% of the structures used in the clustering analysis. The representative structure of this cluster has a core turn region between residues Thr64 and Ile66. The termini are well separated from each other, with the sulfoxide of the Met oriented towards the solvent. The second most populated cluster (9.4%) is in the form of a $\beta$-hairpin with the two aromatic rings being solvent exposed in a $\pi$-stack orientation. This arrangement of the rings has a stabilising effect on the hairpin conformation.

The clustering analysis of the Met60Val and Met60Gln simulations showed similar results with respect to the structures sampled and their relative population, highlighting the structural resemblance of these peptides. Altogether, 1460 and 1358 clusters were classified, with the most populated clusters of the peptides containing 26.1% and 27.8% of the structures, respectively. These structures exhibited the $\beta$-hairpin motif with a core turn region between residues Thr64 and Ile66. Interestingly, the most populated clusters of the Met60Val and Met60Gln mutated peptides exhibited very similar structures to the most populated cluster obtained from the wild-type apoC-II(60-70) simulation. The higher population of the mutated systems compared to the wild-type peptide reflects higher conformational flexibility of these systems which was observed in the secondary structure analysis.


<table>
<thead>
<tr>
<th>Number of clusters</th>
<th>Structure and population of the two most populated clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild-type (1259)</td>
<td><img src="image" alt="Image of Wild-type" /> 50% 5.2%</td>
</tr>
<tr>
<td>Oxi-Met (1417)</td>
<td><img src="image" alt="Image of Oxi-Met" /> 19.4% 9.4%</td>
</tr>
<tr>
<td>Met60Val (1460)</td>
<td><img src="image" alt="Image of Met60Val" /> 26.1% 13.6%</td>
</tr>
<tr>
<td>Met60Gln (1358)</td>
<td><img src="image" alt="Image of Met60Gln" /> 27.8% 15.3%</td>
</tr>
</tbody>
</table>

**Figure 6.3** The number of clusters identified using an RMSD cutoff of 0.15 nm for each individual simulation is shown in the left of the figure. Structural representation of the two most populated clusters and their population are presented on the right.
In the second most highly populated cluster, the structures exhibit slightly distorted hairpin conformations, mostly differing in the arrangement of the aromatic rings. The significance of the relative orientation of the aromatic side-chains in individual peptides is discussed in Section 6.4.6 of this chapter.

The formation of stable β-hairpin structure was recently observed in a fibril-forming 12 amino acid sequence denoted as LSFD [338]. The core hairpin turn in this work consisted of a Gly-Ala sequence. We note that in the simulated apoC-II(60-70) peptides, the location of the Gly is central in the three residue turn region (Thr64, Gly65, Ile66). This suggests a nascent formation of turns and bends at the Gly residue, leading to an overall U-shaped topology (strand-turn-strand) of the backbone which facilitates the formation of intramolecular sheets.

Overall, using clustering methods we were able to identify the specific features in the conformations of the most populated clusters of the wild-type and mutated peptides. The fibril-inhibiting oxidised Met exhibited structure which showed a tendency for a marked increase in separation of the termini. The average C\textsubscript{\alpha} distance between the N- and C-termini of structures contained in the most populated cluster of apoC-II(60-70) with oxidised Met was 1.15 nm. The Met60Val and Met60Gln mutations formed a very stable β-hairpin motif, similar to the wild-type peptide, with the two termini being in very close proximity to one another. The average inter-termini distance of structures contained in the most populated clusters of the wild-type, Met60Val and Met60Gln mutated apoC-II(60-70) peptides was 0.5 nm, 0.6 nm and 0.64 nm, respectively. The compact structures formed by the two mutations are stabilised mainly by hydrophobic interactions which could provide favourable conditions for peptide association.

### 6.4.3 Root mean square deviation

The root mean square deviation (RMSD) was calculated between the representative structures of the most populated clusters of each simulation, presented in Figure 6.3. A rotational and translational fit was performed on the full peptide when compared to the reference structure.

The initial RMSD calculations were performed on the Met60Val and Met60Gln peptides with the wild-type apoC-II(60-70) conformation as the reference structure. An RMSD of 2.2 Å and 3.7 Å respectively was found. RMSD calculations were also performed
on the most populated structure of the oxi-Met peptide (reference structure), in comparison with the wild-type and mutated peptides. The RMSD results consisted of wild-type = 5.6 Å, Met60Val = 5.7 Å and Met60Gln = 5.4 Å. The high RMSD values indicated significant structural differences between the oxi-Met and the wild-type as well as the mutated structures. The lower RMSD result indicated a relatively good agreement in the geometry of the most populated structure of the Val and Gln mutated peptides with the wild-type conformation.

Overall, the RMSD calculations of each apoC-II(60-70) peptide gave quantitative measurements of the structural differences between the highly sampled conformations. In particular, RMSD indicated differences in conformations between the fibril-inhibiting (oxi-Met) and fibril-forming (wild-type) peptides, and similarity between the wild-type and mutated peptides.

6.4.4 Radius of gyration

The radius of gyration (Rg) was calculated for each system to give an indication of the distribution of atoms relative to the peptide’s centre of mass. The moving average of the Rg over the entire trajectory of every system is presented in Figure 6.4.

The Met60Val and Met60Gln exhibited similar trend in Rg fluctuations, when compared to the wild-type simulation. All three systems displayed values of Rg between 0.55 nm to 0.65 nm. The similarity in the Rg results suggests analogous overall shape of these systems, regardless of the mutations they contain.

In contrast, the oxi-Met simulation demonstrated several peaks of high Rg, up to 0.75 nm, indicating the presence of less compact structures, most likely due to increased flexibility. These results are in good correlation with the structure evolution plot of oxidised apoC-II(60-70) (Figure 6.2b). The peaks in Rg correspond to the increase in separation between the N- and C-terminus, resulting from the formation of turn structure between residues oxi-Met60 to Tyr63, and coil conformation from Thr68 to Gln70.
CHAPTER 6. Investigation on the effects of mutation of the amyloidogenic apoC-II(60-70) peptide

Figure 6.4 Radius of gyration (nm) of each simulation, represented by a moving average of 5 ns; a) Wild-type; b) oxi-Met; c) Met60Val; d) Met60Gln.
6.4.5 Side-chain interactions

The interactions between the residues were examined by constructing a residue contact map for the wild-type and mutated systems. The contact maps shown in Figure 6.5 present the pairwise distances between all atoms of individual residues, averaged over the structures in the most populated clusters of wild-type, oxidised, Met60Val and Met60Gln peptides.

The contact maps of the wild-type, Met60Val and Met60Gln (Figure 6.5a, c, d, respectively) show similar interaction distances between the residues. The formation of contacts between residues of the opposing ends of the chain, in particular, the last two residues of the N- (60 and 61) and C-termini (69 and 70) can be seen. For example, the contacts of Ser61 with Phe67, Asp69 and Gln70, are in the range of 0.2 nm to 0.25 nm. These distances highlight the contacts that occur between the peptide ends in β-hairpin conformations.

The contact map from the oxi-Met simulation exhibited different results to those described above. The lack of interaction between the terminal residues is indicated by the blue colour, signifying long contact distance (0.5 nm or higher). However, close contacts have formed between the oxi-Met60 with Thr62 and Tyr63, and between Ser61 and Tyr63. These contacts were not present in the wild-type, Met60Val and Met60Gln peptides. The formation of close contacts with neighbouring residues and lack of interaction between the termini residues indicated that these interactions weaken the overall attraction between the strands, causing the peptide to adopt relatively extended conformations.

The interaction of oxi-Met with its neighbouring residues was investigated by hydrogen bond analysis recently. In order to obtain an insight into the inhibition effects of the oxidised Methionine of apoC-II(60-70), Hung et al. examined the H-bonding interactions between the side-chain of the oxi-Met60 and the remainder of the peptide during the simulation described in reference [285]. Two persistent hydrogen bonds were found, between the S=O of the oxi-Met and the backbone of Ser61 and Thr62, while interactions with the other residues were infrequent.
Figure 6.5 Contact maps representing the average contact distance between all atoms of each residue for the structures contained in the most populated cluster of apoC-II(60-70) peptide; a) Wild-type; b) oxi-Met; c) Met60Val; and d) Met60Gln. A contact is defined by a distance of < 0.5 nm.
Aromatic side-chain orientation

The role of the aromatic rings in fibril formation was discovered to be significant, specifically in relation to their presence in fibril forming regions and their contribution to fibril stability [130]. Site-directed modification performed on short amyloidogenic fragments revealed that aromatic residues play a crucial part in the fibrillation process [131]. In the case of the IAPP motif mutation of Phe to Ala there was significant reduction in the peptide’s potential for fibril formation [342]. The presence of two aromatic rings (Tyr63 and Phe67) in the apoC-II(60-70) peptide gives reason to suggest their possible role in aggregation. The orientation of their side-chains is of significant interest, as distinct differences were recently identified between fibril forming and inhibiting conditions [285].

In order to reveal the effect of mutation on the aromatic side-chain orientation the relative position of the aromatic side-chains with respect to the peptide backbone structure was determined. The average angle between the Cα-Cγ vector of Tyr63 and the Cα-Cγ vector of Phe67 was calculated, during the equilibrium part of the simulation of the Met60Val and Met60Gln peptides. Angles of less than 90° indicate that the two side-chains lie on the same “face-side” of the peptide structure, while angles greater than 90° indicate that the side-chains are on the opposite faces of the hairpin. Results from our recent work [285] showed that the peptide under fibril-favouring conditions adopts structures where Tyr63 and Phe67 prefer the opposite “side” of the hairpin structure, whereas the structures obtained under fibril-inhibiting conditions (oxi-Met) have the aromatic rings on the same side of the hairpin. The histograms and representative structures obtained from the Met60Val and Met60Gln simulations are shown in Figure 6.6. The angles determined for oxi-Met and wild-type apoC-II(60-70) from our previous work [285], are also presented for comparison.

The angles adopted by the aromatic side-chains of Met60Val and Met60Gln peptides lie approximately between 130° and 150°, similar to the wild-type peptide, with the majority of the structures containing the aromatic rings located on opposite faces of the hairpin structure. The similar orientation of the Tyr63 and Phe67 side-chains between the mutated and native peptide suggests favourable arrangement for fibril formation. However, the broader distribution of angles for the mutated systems suggests possible differences in kinetics and/or fibril morphology for these peptides.
CHAPTER 6. Investigation on the effects of mutation of the amyloidogenic apoC-II(60-70) peptide

Figure 6.6 Histograms of Tyr63 ring orientation with respect to Phe67 (x-axis), obtained from total simulated trajectories for oxidised Methionine [285] (red), Met60Val (green), Met60Gln (blue) and wild-type (black) peptides. Y-axis indicates the percentage of structures for a specific orientation angle range exhibited during the trajectory. Forty bins were used over the range 0 – 180°. Each data point represents the percentage (%) of structures with orientation angles between x and x + 4.5°. Angle < 90° indicates both rings are on one side of the hairpin, while > 90° indicates rings exist on opposite sides. Structures illustrating the relative ring orientations are represented in insets.
6.4.7 Solvation properties

The solvation properties and conformational changes due to solvent interactions with the peptide are important, as shown in a recent study by Wei and Shea on the effect of solvents on the conformations of the Aβ peptide [339]. This study highlighted the influence of solvent on the structured states of amyloidogenic peptides that may result in fibril formation or inhibition. Their simulations revealed that the Aβ(25-35) peptide preferentially populated a helical structure in an apolar organic solvent (postulated to be fibril inhibiting), while in pure water, the peptide adopted collapsed coil and β-hairpin conformations which did form fibrils. Additionally, Fernandez and Scheraga [343] have noted that proteins that aggregate readily tend to have a significant number of backbone H-bond donors/acceptors exposed to the solvent, available for protein interaction. Therefore the exposure to solvent for further peptide association is important.

The average solvent accessible surface area (SASA) and average number of water molecules found in the first hydration shell of each peptide were calculated and are presented in Figure 6.7. The averages were determined from the structures contained in the most populated cluster of each simulation. The SASA was calculated for each system using a probe radius of 1.4 Å and Lennard-Jones hard-shell radii for each atom to define the surface. The number of water molecules present in the first hydration shell, that is within a 3 Å distance of each residue, was also calculated.

Interestingly, there is only a small difference in the total SASA of the structures within the most populated clusters of each system, wild-type = 11.41 nm$^2$, oxi-Met = 11.34 nm$^2$ and Met60Val and Met60Gln = 11.23 nm$^2$, respectively. However, the SASA per residue for the fibril-inhibiting mutation (oxi-Met) is very different to the SASA of the wild-type, Met60Val and Met60Gln peptides.

In the oxidised system, the SASA per residue of the oxi-Met60 and Gln70 residues are almost equal in value, while Gln70 has on average two more water molecules within a 3 Å shell compared to oxi-Met (Figure 6.7b). This is a result of the side-chain interaction of oxi-Met with the Ser61 and Thr62, while Gln70 is exposed for full solvation. The destabilising effect of oxidation of Met60 on the interaction between the strands contributes to a higher SASA in this region compared to the wild-type and mutant systems where attractive interaction between the strands minimises their exposure to the solvent. In contrast, the SASA of the core turn region (Thr64-Ile66) of the oxidised peptide is lower compared to the mutated
peptides, which in turn have similar SASA to the wild-type apoC-II(60-70). The hydrophobic core between residues Thr64, Gly65 and Ile66 is retained within the structures of the most populated cluster of the oxidised system, which results in the observed decrease in SASA of these residues and reduced number of surrounding water molecules (Figure 6.7a).

Our simulations indicate that the residues comprising the $\beta$-turn region of the mutants, Thr64, Gly65 and Ile66, are more exposed to the water than those of the oxidised peptide (Figure 6.7a). This suggests that the $\beta$-turn in Met60Val and Met60Gln may be responsible for initiating aggregation. Furthermore, the similarities seen in Figure 6.7a for Val and Gln mutated fragments with the fibril-forming wild-type peptide [285], suggest that aggregation could be a nascent property for these apoC-II(60-70) peptides.

The low solvent exposure in the oxidised system of the hydrophobic core region (Thr64, Gly65, Ile66) may inhibit interactions with “like” peptides. In contrast, in the wild-type and mutant systems, the exposure of the hydrophobic region is higher, so the interactions between “hydrophobic patches” from different peptides can drive their aggregation. Similar hydration/dehydration effects on the conformational preferences of A$\beta$(25-35) in pure water were observed by Wei and Shea [339].
CHAPTER 6. Investigation on the effects of mutation of the amyloidogenic apoC-II(60-70) peptide

Figure 6.7 Average values calculated per residue over the structures contained in the most populated cluster of the oxidised Methionine (red), Met60Val (green), Met60Gln (blue) and wild-type (black) simulation. a) Solvent accessible surface area (nm$^2$); b) The average number of water molecules within a 3 Å shell of each residue.
6.5 Conclusions

The investigation of the conformational preferences adopted by monomeric proteins provides important insight into the possible features that influence their fibril-forming or inhibiting propensities. Computational techniques are useful for such studies, because these features can be investigated atomistically under controlled conditions. The recent discovery of the variation in fibrillation kinetics and morphology between variants of the full-length [281] and shorter segments (56-76 [284] and 60-70 [285]) of the apoC-II protein is of great interest and motivated additional detailed investigation of the mechanisms involved.

In this chapter, the effect of single-point mutation at Met60 to hydrophobic Val (Met60Val) and hydrophilic Gln (Met60Gln) on the dynamics and structures of the apoC-II(60-70) peptide were investigated. The results were compared with the wild-type and oxidised apoC-II(60-70) peptide [285], as the mutants were chosen to mimic these systems.

Structural features and properties, such as the formation of a β-hairpin, solvent accessible surface area and relative orientation of the aromatic side-chains showed similarity between the Met60Val and Met60Gln peptides. Furthermore, the systems exhibited features identified as fibril-favouring for the native peptide [285].

Based on the side-chain interaction and orientation analysis performed on the two mutations Met60Val and Met60Gln, some important differences were also identified, such as the broader distribution of angles observed between the aromatic Tyr63 and Phe67 residues. Therefore, it is reasonable to suggest that the mutants may form fibrils, although likely with different kinetics and/or fibrillar morphology. This hypothesis is in agreement with the findings of a recent study performed by Legge et al. [284]. It was shown, both experimentally and computationally, that Met60Val and Met60Gln mutations of the apoC-II(56-76) peptide formed fibrils similar to the wild-type aggregates, however fibril formation proceeded at different rates between the different peptides. The experimental studies to test our findings on the apoC-II(60-70) peptide are currently underway.

From entropic considerations, the most probable candidates to initiate further fibril growth are those with pre-formed structure, such as β-hairpins. Comparison of the mutant systems with the oxidised system revealed insights into mechanisms of aggregation. All systems exhibited strong hydrophobic core regions, however, the mutant and wild-type systems also displayed increased solvent exposure of this region, compared to the oxidised...
system. From this we can infer that the β-turn region of the peptides (Thr64, Gly65, Ile66) is a segment responsible for initiating the aggregation process, as fibril formation is mainly driven by availability of the backbone atoms for hydrogen bonding and hydrophobic interactions with nearby molecules.

Overall, our results show that MD simulations can be used to identify conformational preferences responsible for aggregation as well as a qualitative predictor of the peptide fibrillogenicity.
Chapter 7

7. Lipid concentration effects on the amyloidogenic apoC-II(60-70) peptide

7.1 Overview

Recent experimental studies have shown that micellar and sub-micellar concentrations of short-chained phospholipids have different effects on the propensities for fibril formation of the native and shorter segments of the apoC-II protein [141, 142]. This finding highlighted the need for improved understanding of the lipid-peptide interactions and their influence on the mechanisms of fibril formation.

In this chapter, explicit solvent molecular dynamics simulations were implemented to investigate the effect of phospholipid concentration on the conformation of the amyloidogenic peptide apoC-II(60-70). The dipentanoylphosphatidylcholine (D5PC) lipid was used, with concentrations of 1 to 6 lipids per peptide. To investigate the effects of different starting peptide conformations, calculations were initiated from two distinct structures, random coil and α-helix. The umbrella sampling technique was applied to determine the effect of lipid concentration on the free energies of apoC-II(60-70) as a function of the peptide interterminus center-of-mass separation (peptide stretching). The molecular mechanisms of lipid effects on the peptide conformation have been identified and are discussed in this chapter.
7.2 Introduction

Amyloid diseases are unified by a characteristic presence of fibrous deposits of misfolded proteins, however recent studies have revealed a number of other pathophysiological features which appear to be common to many amyloid related conditions. One such common feature is the involvement of lipids at various stages of the evolution of the disease. A detailed literature review on the involvement of lipids in amyloid diseases is presented in Chapter 1. One example in particular for Alzheimer’s disease was the identification of the apolipoprotein E (apoE) genotype and its connection with lipid metabolism as a major genetic risk factor for the disease [344, 345]. Another member of the very-low-density lipoproteins associated with amyloid diseases is apolipoprotein C-II (apoC-II), which exhibits different fibrillation propensities and kinetics depending on the concentration of lipid in its environment [142, 144]. A detailed description of the structure and function of apoC-II is presented in Chapter 3 of this thesis.

The solution structure of apoC-II is predominantly $\alpha$-helical in the presence of micellar concentrations of lipids. However, under lipid-poor conditions apoC-II readily aggregates to form homogeneous fibrils [142, 144, 277, 346]. In a lipid-free environment the structural stability of the apolipoprotein is compromised and the molecule adopts random coiled structures which are thought to contribute to its propensity to aggregate into amyloid fibrils. The presence of macro-molecular crowding agents [280], auxiliary proteins [279] and the habitation of phospholipids below and above the critical micelle concentrations (CMC) in solution alter the kinetics of apoC-II fibril formation and fibril morphology [144]. In recent work by Griffin et al. it was shown that sub-micellar concentrations of short-chain phospholipids increase the rate of apoC-II fibril formation in a manner dependent on lipid chain length and concentration. However, high micellar concentrations of phospholipids completely inhibited amyloid formation [142]. At lower concentration of soluble phospholipid complexes, fibril formation by apoC-II was only partially inhibited. Using electron microscopy (EM) and circular dichroism (CD) spectroscopy it was found that phospholipid complexes change the structural architecture of mature fibrils and generate new fibril morphologies [142].
In a more recent publication, experimental and computational techniques were used to study the effects of dihexanoylphosphatidylcholine (D6PC) and dipentanoylphosphatidylcholine (D5PC) on fibril formation and dimerisation by the tryptic peptide apoC-II(60-70) [141]. Thioflavin T fluorescence studies showed that sub-micellar levels of these short chain phospholipids strongly inhibited amyloid fibril formation by the 11 residue apoC-II(60-70) peptide. At the same time, sedimentation equilibrium analysis indicated the presence of soluble oligomeric complexes. Molecular dynamics simulations coupled with umbrella sampling were performed to determine the dimerisation free energies of a number of β-stranded and random coil dimer complexes, both in the presence and absence of lipids. The simulations indicated that in contrast to their inhibitory effects on fibril formation, short-chain phospholipids promote the formation and stabilisation of dimers by enhancing inter-subunit hydrophobic interactions [141]. Based on these results, the authors proposed the mechanism by which peptide bound lipids inhibit fibrillisation is by trapping the dimers (and other oligomeric species) in arbitrary conformations, including fibril-disfavouring ones, reducing their likelihood to dissociate and re-associate into conformations more prone to fibril nucleation and growth. This is consistent with the computational study by Jiang et al. in which the replica exchange method (REMD) was utilised to investigate the influence of a single lipid molecule, dioleoylphosphatidylcholine (DOPC) on the aggregation of the hIAPP(20-29) peptide [347]. It was found from the simulations that the lipid exerted a stabilising effect on the ordered β-sheet structure via hydrophobic interactions at positions Phe23-Pro25. They suggested that the lipid can act as a catalyst in hIAPP(20-29) peptide self-assembly.

Most computational investigations conducted so far have looked at the effects of lipid bilayers or membranes on the structure and aggregation propensities of peptides, as presented in the literature review in Chapter 1. To the best of our knowledge the only studies that have investigated the effects of various concentrations of free lipids on the structure and dynamics of proteins are those mentioned above. However, in these papers the mechanisms of interactions and effects of free lipids at various concentrations on the structure and dynamics of amyloidogenic peptides are not described in detail.

In this chapter, atomistic molecular dynamics (MD) simulations and free energy (ΔG) calculations were applied to investigate the conformational changes of apoC-II(60-70) in the presence of D5PC molecules at different concentrations in solution. The interactions between
solvated lipids and peptides, and consequent effects on the peptide structure that may take place in bulk solution at sub-CMC conditions were only considered. Specifically, we have concentrated on the conformational changes in the peptide caused by association with lipids to gain an insight into the atomistic interactions between the solutes that may discriminate between the fibril forming or inhibiting effects.

7.3 Computational Details

Investigation of the effects of lipids at various concentrations on the structure and dynamics of the apoC-II(60-70) peptide was performed by employing equilibrium molecular dynamics (MD) and umbrella sampling simulations [200]. The molecular dynamics simulations were performed in NPT ensemble using the Gromacs 3.3 [209] simulation package. The united-atom Gromos force field was implemented with the parameter set 43A1 [174]. Force field parameters for the D5PC lipid molecules were derived from those of Berger et al. [348]. The time step for all the simulations was set to 2 fs. The electrostatic and Lennard-Jones interactions were calculated with a cutoff of 10 Å. The Particle Mesh Ewald (PME) [294] method was used to account for the long-range electrostatic interactions. All bond lengths were constrained to their equilibrium values using the LINCS [193] algorithm.

The effects of various lipid concentrations were investigated by performing systematic simulations with increasing lipid to peptide ratio from 1:1 to 6:1. These concentrations correspond to the sub-micellar concentration of D5PC. Two different starting conformations of apoC-II(60-70) were selected for the MD simulations at each concentration, coiled and helical, in order to enhance conformational sampling and to investigate the interactions of lipids with different structural elements. The residue sequence of the apoC-II(60-70) peptide, as well as the starting conformations used in the simulations with D5PC lipid are shown in Figure 7.1.

Each peptide-lipid system was enclosed in a periodic box of selected size, ranging from 42.5 Å × 42.5 Å × 42.5 Å for the smallest lipid to peptide ratio of 1:1 to 44 Å × 44 Å × 44 Å for the 6:1 ratio simulations. Each simulation box was explicitly solvated using the SPC water model, corresponding to a water density of ~1.0 g/cm$^3$. The simulation box size was specifically assigned to maintain approximately the same number of water molecules so the
CHAPTER 7. Lipid concentration effects on the amyloidogenic apoC-II(60-70) peptide

effect of lipid concentration on the peptide conformation can be evaluated. To reduce initial lipid-peptide interactions, each lipid was placed at a random location in the periodic cell with a minimum distance of 6 Å from the peptide.

Figure 7.1 Structural representation of D5PC lipid molecule, where the carbon atoms are represented in cyan, hydrogen in red and the phosphate in green. The apoC-II(60-70) starting conformations, coil and helix are also shown together with the residue sequence of the peptide.

Each system was energy minimised to remove steric clashes using the steepest descent algorithm with convergence criterion of 20 kJ/mol/nm. To allow the solvent to relax around the protein, equilibration of 1 ns was performed at constant pressure (NPT) with the protein restrained. Constant temperature conditions at 300 K and pressure of 1 bar were achieved by coupling the system using Berendsen thermostat and barostat [188]. Data collection was accumulated in an NPT ensemble for each simulation for the total time presented in Table 7.1a.

The umbrella sampling [200] methodology was applied to calculate the free energy ($\Delta G$) of the peptide as a function of centre-of-mass (COM) separation of the $C_\alpha$ atoms of the N- and C-termini at each lipid concentration, effectively sampling the peptide stretching. Umbrella sampling simulations of the peptide in the lipid-free environment were also performed to allow a comparison to be made between the free energies of the peptide extension in different environments. The separation of the termini between 0 and 2 nm was maintained in each window by Hookean functions with force constants of 20 kcal/mol/Å$^2$. 

<table>
<thead>
<tr>
<th>60</th>
<th>61</th>
<th>62</th>
<th>63</th>
<th>64</th>
<th>65</th>
<th>66</th>
<th>67</th>
<th>68</th>
<th>69</th>
<th>70</th>
</tr>
</thead>
<tbody>
<tr>
<td>Met</td>
<td>Ser</td>
<td>Thr</td>
<td>Tyr</td>
<td>Thr</td>
<td>Gly</td>
<td>Ile</td>
<td>Phe</td>
<td>Thr</td>
<td>Asp</td>
<td>Gln</td>
</tr>
</tbody>
</table>
Adjacent windows were separated by 0.05 nm. In total, 40 windows were simulated for 10 ns each. The Weighted Histogram Analysis Method (WHAM) [349] was subsequently applied to obtain the unbiased free energy profiles (potentials of mean force, PMF). The total simulation times accumulated using the umbrella sampling method are listed in Table 7.1b.

All the simulations in this work are labelled based on the lipid concentration (ratio of lipid to peptide) and type of starting structure with the system codes listed in Table 7.1. For example, for the simulations having 1 peptide starting from a coiled conformation with 1 lipid concentration, the system was labelled 1lc, while the simulation with 1 lipid and with helical starting conformation for the peptide was labelled 1lh. The simulations with ratio 2:1 were labelled, 2lc and 2lh, and so forth. For the 6:1 lipid to peptide system, starting from a helical conformation 2 separate simulations were performed, labelled 6lh and 6lh2, because of the results we obtained in the first simulation; this is discussed in detail in the Results and Discussions section of this chapter. In the second simulation 6lh2, only the seed for the random number generator of the initial velocities was changed, while the starting conformation of the system was the same as in 6lh.

To aid in the presentation and interpretation of the results each system with different lipid concentration was assigned a colour. In particular, the radial distribution function (RDF) of water, the lipid-peptide interaction histograms and the umbrella sampling free energy profiles are shown using the following colouring scheme. The lipid free simulation is represented in red, 1:1 ratio simulations in green, 2:1 ratio in blue, 3:1 ratio in magenta, 4:1 ratio in cyan, 5:1 ratio in yellow and the 6:1 ratio in black.

In the Results and Discussions section of this chapter we firstly present the results obtained from the classical molecular dynamics simulations, such as detailed analysis on the lipid-peptide interactions and their effect on the evolution of the secondary structure. The analysis of the MD simulations is followed by discussion of the results obtained from the umbrella sampling simulations, such as the free energy profiles at various lipid concentrations. The implication of our results on the effects of lipid concentration on the structure, dynamics and aggregation of the apoC-II(60-70) peptide are also discussed.
CHAPTER 7. Lipid concentration effects on the amyloidogenic apoC-II(60-70) peptide

Table 7.1 The nomenclature for the individual system, as well as their simulated times using: a) Molecular dynamics and b) Umbrella sampling method. The two digits in the label describe the ratio of lipid to peptide, and the letters c and h describe coil and helix starting structure, respectively.

<table>
<thead>
<tr>
<th>System</th>
<th>Simulation Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1lc/1lh</td>
<td>550 ns</td>
</tr>
<tr>
<td>2lc/2lh</td>
<td>550 ns</td>
</tr>
<tr>
<td>3lc/3lh</td>
<td>550 ns</td>
</tr>
<tr>
<td>4lc/4lh</td>
<td>550 ns</td>
</tr>
<tr>
<td>5lc/5lh</td>
<td>550 ns</td>
</tr>
<tr>
<td>6lc</td>
<td>550 ns</td>
</tr>
<tr>
<td>6lh</td>
<td>300 ns / 300 ns</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>System</th>
<th>Simulation Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>ambient</td>
<td>10 ns</td>
</tr>
<tr>
<td>1lc</td>
<td>10 ns</td>
</tr>
<tr>
<td>2lc</td>
<td>10 ns</td>
</tr>
<tr>
<td>3lc</td>
<td>10 ns</td>
</tr>
<tr>
<td>4lc</td>
<td>10 ns</td>
</tr>
<tr>
<td>5lc</td>
<td>10 ns</td>
</tr>
<tr>
<td>6lc</td>
<td>10 ns</td>
</tr>
</tbody>
</table>

7.4 Results and Discussions

7.4.1 Molecular dynamics: Lipid-peptide interactions

In the Introduction section of this chapter it was highlighted that various concentrations of lipids can have diverse effects on the structure of proteins and their propensity to form fibrils, thus detailed study of the lipid-peptide interactions is required. In this work, we were specifically interested in the type of interaction occurring between the apoC-II(60-70) peptide and the lipid polar head-group or hydrophobic alkyl tail of D5PC. To determine the general distribution of lipid components around the hydrophobic and hydrophilic regions of the peptide, radial distribution functions (RDF) were calculated. Typical RDFs observed of the lipid head/s and tail/s with respect to the hydrophobic (Tyr63, Gly65, Ile66, Phe67) and hydrophilic (Met60, Ser61, Thr62, Thr64, Thr68, Asp69, Gln70) residues of the apoC-II(60-70) peptide are shown in Figure 7.2. The RDFs were calculated for all systems and are presented in the Appendix of this thesis.
The RDF profiles exhibited higher density of lipid tails in the proximity of the hydrophobic residues, whereas the polar lipid-heads were mostly surrounding the hydrophilic residues of the apoC-II(60-70) peptide. Specifically, the head-philic groups showed a peak at ~2 Å indicative of hydrogen bonding, while the tail-phobic groups showed a broader peak at ~5 Å indicative of van der Waals (vdW) interactions, including hydrophobic association. These results were not surprising as alkyl tails favour interactions with hydrophobic surfaces, while the lipid-heads favour hydrophilic surfaces. This trend was common among all systems indicating vdW contacts between the lipid and the peptide, except for 1lc where higher density of lipid-heads was observed in the region of hydrophilic residues. Interestingly in some systems, in particular those with 3 or more lipids per peptide, high density of lipid-heads can be seen in the proximity of hydrophobic residues also, which could suggest a micelle-like formation where the hydrophobic tails are buried in the peptide while the hydrophilic heads are exposed to the solvent, illustrated in Figure 7.2. As expected, the lowest presence of lipid tails was seen in the proximity of the hydrophilic residues.

![Figure 7.2](image.png)

**Figure 7.2** Typical radial distribution functions of the lipid head/s and tail/s with respect to hydrophobic and hydrophilic residues of the apoC-II(60-70) peptide. The RDFs for the 4lh system are shown as an example, while for all the other systems the RDF plots are presented in the Appendix of the thesis.
 CHAPTER 7. Lipid concentration effects on the amyloidogenic apoC-II(60-70) peptide

Figure 7.3 Schematic of the micellar-like formation of lipids surrounding the peptide. The hydrophobic residues are represented in blue, hydrophilic in red and the lipid in black.

A more quantitative analysis was performed on the total number of contacts and their stabilities, formed between the lipid components, head and tail, with each residue of apoC-II(60-70). A contact was defined where the distance between any pair of atoms of the peptide’s residue and the respective lipid component, head or tail, was less than 0.4 nm. In Figure 7.4 the data for the systems which exhibited different behaviours, I1c and 6lh is shown. The results for all systems are presented in the Appendix of the thesis.

Close inspection of the plotted histograms in Figure 7.4 showed that a higher number of contacts were made by the lipid head group/s, compared to the tails, with all the residues of apoC-II(60-70). This result was expected because the head group is twice as large as the tail, 20 and 10 atoms, respectively. However, there was a significant difference in the stability of the contacts made by the individual lipid components. In particular for the I1c system, the contacts made by the lipid tail with the hydrophobic region from Tyr63 to Phe67 were more stable than the contacts made by the lipid head. In contrast, higher contact stabilities were observed between the lipid head and the hydrophilic Met60, Ser61, Asp69 and Gln70 residues. This is shown by the longer periods of interactions experienced between these groups (top histogram of Figure 7.4).

In the 6lh simulation, the highest numbers of contacts were made by the whole lipids with the aromatic residues, Tyr63 and Phe67. Significant contacts were also made by the lipid head with the hydrophilic Thr68 to Gln70 residues. In contrast to the I1c system, there was less difference in the contact stabilities between the individual lipid groups and the residues.
This result is expected, due to the higher number of lipids that are interacting with the peptide. However, this result also highlights that stable, continuous interactions were made by the whole lipids with the hydrophobic region of apoC-II(60-70) between residues Tyr63 and Phe67, which could indicate micelle-like formations at this concentration, similar to the illustration in Figure 7.3.

In order to identify the favourable lipid interaction sites on the peptide and the exposure to water at various lipid concentrations, the average solvent accessible surface (SAS) area per residue was calculated for each simulation. A probe radius of 1.4 Å and Lennard-Jones hard-shell radii for each atom were used to define the SASA. In Figure 7.5 the average SASA and the corresponding error per residue for each molecular dynamics simulation is presented. To enable comparison to be made with the SASA of the peptide in a lipid-free environment, calculations over the full trajectory of the data collected for the wild-type apoC-II(60-70) (Chapter 6) are also presented in grey of Figure 7.5.

We observed distinct reduction in SASA between the ambient and lipid-environment systems in the proximity of the two aromatic residues, Tyr63 and Phe67, present in apoC-II(60-70). This trend was seen in all systems, regardless of the starting conformation or lipid concentration. In particular, the lowest SASA was seen between residues Thr62 and Tyr63 (1st half of the peptide, strand 1) and Ile66 and Phe67 (2nd half of the peptide, strand 2) of apoC-II(60-70). This trend was observed in all simulations, although the difference in SASA of the peptide in lipid-free and lipid-environment was higher with increasing lipid-concentration, further highlighting the favourable interactions of the lipids with the aromatic and their neighbouring residues.

Furthermore, there were significantly smaller differences in the SASA between the first half of the peptide (strand 1) compared to the second half (strand 2). In particular, only a small difference in SASA was observed in the region of Met60 and Ser61, at all lipid concentrations. This finding is not surprising as there are more hydrophobic residues in strand 2 (Gly65, Ile66 and Phe67), compared to the only hydrophobic residue Tyr63 in strand 1. Additionally, Asp69 and Gln70 in strand 2 have higher hydrophilicity than Met60 and Ser61 in strand 1, thus the lipid head groups are more strongly drawn to interact with strand 2 of apoC-II(60-70). Higher number of contacts between the whole lipids and the aromatic residues, as well as the lipid head groups with Asp69 and Gln70, were also observed in the contact histogram plots of Figure 7.4 and the Appendix.
CHAPTER 7. Lipid concentration effects on the amyloidogenic apoC-II(60-70) peptide

Figure 7.4 Histograms representing the contacts formed between the lipid head and tail with individual residues of apoC-II(60-70) for systems 1lc and 6lh. Percentages of the times in contact are shown in the red/blue histograms, while the percentages of total number of contacts are shown in the grey/green histograms of each system.
CHAPTER 7. Lipid concentration effects on the amyloidogenic apoC-II(60-70) peptide

Residues
Figure 7.5 Average solvent accessible surface area (SASA) per residue and corresponding error.
Aromatic regions play an essential role in binding lipids to various other apolipoproteins. Structural studies performed using NMR of the apoA-I [350], apoE [351] and apoC-I [352] proteins pointed out the importance of aromatic residues in anchoring apolipoproteins to lipid micelles. However, an additional role of the aromatic residues in self-association of these proteins was also identified. In particular, the aromatic stacking in the C-terminal domain of apoA-I plays a key role in protein aggregation [353]. Similarly, the aromatic-rich C-termini of apoE [354] and apoC-I [355], known to be important in lipid binding, are also essential for apolipoprotein aggregation. Burley and Petsko found that aromatic-aromatic interactions play a key role in apolipoprotein aggregation and subsequent fibril stability [356].

In our studies, the results obtained from the SASA and lipid-residue contact analysis suggest frequent and favourable interactions between the lipids and the hydrophobic residues of the peptide. In particular, we found that Tyr63 and Phe67 residues present in the apoC-II(60-70) peptide act as interaction sites with lipids. The lipids binding to aromatic residues can be seen as having an inhibitive effect on fibril formation as these residues become less exposed for potential peptide association.

We proceeded to investigate the number of lipids that are in contact with the peptide at every step of the entire trajectory. A contact was identified when the distance between any atom of the peptide and any atom of a lipid was less than 0.4 nm. The number of interacting lipids at each frame of every system was counted and the percentages of interacting lipids in each trajectory are presented as histograms in Figure 7.6.

The results presented in Figure 7.6 illustrate the relative population of lipids that are in contact with the peptide over the entire trajectory. At low concentrations, such as the lipid:peptide ratios of 1:1, 2:1 and 3:1, we found that all lipids present in the system were in contact with the peptide for at least 83% of the time. The distinct difference in the percentage of interacting lipids in the 3lc and 3lh systems was due to the conformations of the peptides during these simulations (discussed later).

In the simulations of higher lipid concentrations, such as 4:1, 5:1 and 6:1, a broader distribution of the number of interacting lipids was observed. Specifically, 4 to 5 lipids were found to be continuously interacting with the peptide in ~50% of the total simulated time. Due to the small size of the peptide, this number of lipids could be sufficient to fully cover (coat) the peptide.
CHAPTER 7. Lipid concentration effects on the amyloidogenic apoC-II(60-70) peptide

<table>
<thead>
<tr>
<th>Number of lipid – peptide contacts (%)</th>
<th>Number of lipids</th>
</tr>
</thead>
<tbody>
<tr>
<td>94.35</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>98.41</td>
<td>2</td>
</tr>
<tr>
<td>82.78</td>
<td>4</td>
</tr>
<tr>
<td>84.16</td>
<td>6</td>
</tr>
<tr>
<td>83.29</td>
<td>8</td>
</tr>
<tr>
<td>82.78</td>
<td>10</td>
</tr>
<tr>
<td>62.31</td>
<td>12</td>
</tr>
<tr>
<td>83.29</td>
<td>14</td>
</tr>
<tr>
<td>58.49</td>
<td>16</td>
</tr>
<tr>
<td>36.61</td>
<td>18</td>
</tr>
<tr>
<td>4.88</td>
<td>20</td>
</tr>
<tr>
<td>0.54</td>
<td>22</td>
</tr>
<tr>
<td>15.28</td>
<td>24</td>
</tr>
<tr>
<td>0.29</td>
<td>26</td>
</tr>
<tr>
<td>10.83</td>
<td>28</td>
</tr>
<tr>
<td>39.44</td>
<td>30</td>
</tr>
<tr>
<td>49.44</td>
<td>32</td>
</tr>
</tbody>
</table>

Number of lipids: 1, 2, 3, 4
Figure 7.6 Histogram representation of the % of lipids that were in contact with the peptide for every simulation. The bars of the histogram plots are coloured as described in the Computational Details section of this chapter.
CHAPTER 7. Lipid concentration effects on the amyloidogenic apoC-II(60-70) peptide

The density of water around the peptide in the presence of lipids at various concentrations was also investigated. The analysis was performed by calculating the radial distribution functions (RDF) of the water molecules with respect to the peptide. Figure 7.7 shows the RDF profiles of water for the apoC-II(60-70) peptide starting from the coil and helix conformations, in the presence of D5PC lipids at different concentrations. The RDF calculations were performed on the data collected from the last ~100 ns of molecular dynamics simulations of each system, because at these times the system could be considered to be at an equilibrium state. This was monitored by the root mean square deviations (RMSD) of the peptide, where in each simulation the last 100 ns exhibited flattening in RMSD fluctuations (not shown).

Close inspection of Figure 7.7 showed that the first peak of $g(r)$ at ~1.85 nm is slightly higher for the systems with low lipid concentrations, such as those of lipid to peptide ratios of 1:1 and 2:1, indicating higher number of hydrogen bonds are forming between the water and the peptide. Lower level of water solvation was observed for systems with higher lipid concentrations, such as 5:1 and 6:1, as the insets of Figure 7.7 demonstrate. This trend was similar between all the simulations, regardless of the starting conformation of the peptide.

In the simulations starting from the helix conformation the lipid concentration of 1:1, 2:1 and 3:1 lipids per peptide exhibited similar water RDF, and they can be seen overlaying each other. The higher lipid concentration simulations (lipid:peptide ratios of 5:1 and the two separate 6:1 simulations) showed the lowest water solvation. The RDF of the 4:1 system lies distinctly between the lower and higher solvated systems. This number of lipids was found to be persistently in contact with the apoC-II(60-70) peptide at higher concentrations of 4 or more lipids (Figure 7.6).

Overall, the results demonstrated a desolvation effect by the lipids on the conformation of the apoC-II(60-70) peptide. As expected, with increasing lipid concentration we observed a reduction in specific interactions between water and the peptide due to the lipids binding. A recent study on the effect of lipids on the dimerisation of apoC-II(60-70) found that the greater stability of the dimer could be attributed to the higher peptide desolvation rate upon dimerisation [141].
CHAPTER 7. Lipid concentration effects on the amyloidogenic apoC-II(60-70) peptide

Figure 7.7 Water radial distribution functions (RDF) with respect to apoC-II(60-70) peptide at various lipid concentrations; a) Water RDF profiles of the simulations starting from the coil conformation; b) Water RDF profiles of the simulations starting from the helix conformation. The region of the first RDF peak is shown as an inset. The RDF profiles are coloured as described in the Computational Details of this chapter.

161
7.4.2 Molecular dynamics: Secondary structure evolution

To identify the key conformational changes the peptide experiences under the influence of lipid at various concentrations the STRIDE algorithm [210] was applied to analyse the secondary structure evolution of each system. Typical structural models of the lipid-peptide complexes of selected systems are shown in Figure 7.8. The secondary structure evolution plots of all systems are presented in Figure 7.9.

Inspection of Figure 7.9 demonstrates significant changes in secondary structure of the peptide as the lipid concentration is increased, regardless of the starting structure. The simulations starting from a coiled conformation exhibited reduction in β-sheet elements (yellow) with increasing lipid concentration. The 1:1 system (1lc) forms a β-hairpin structure represented by a turn at residues Thr64 and Gly65 and strong electrostatic interactions between the two strands of the hairpin, residues Ser61-Tyr63 (strand 1) and Ile66-Thr68 (strand 2) forming a β-sheet. We postulated this type of structure to be a favourable conformation for protein association in our recent study on the effects of mutations on the wild-type apoC-II(60-70) presented in Chapter 6. The formation of β-hairpin structure in short amyloidogenic peptides has been observed in molecular dynamics studies of various other proteins (summarised in Chapter 1 and 6). The lipid was in contact with the peptide for the entire simulation, without having a significant effect on its structure, as seen in the structural model shown in Figure 7.8.

As the lipid concentration increased from 2:1 to 5:1 lipids per peptide (2lc, 3lc, 4lc and 5lc, respectively), significant fluctuations in secondary structure were observed and the β-hairpin formation seen in the initial stages of the simulations was soon lost. A reduction in the side-chain interactions between residues Ser61-Tyr63 and Ile66-Thr68 resulted in a frequent extension in the overall shape of the peptide. In Figure 7.8a structural model typical of the extended conformations sampled in the 2lc simulation is shown. The majority of structures sampled at higher lipid concentration were mostly governed by turn and coil structural elements. Interestingly, at lipid concentration of 6:1 (system 6lc) there were fewer fluctuations observed in the secondary structure, with the peptide forming a stable turn region between residues Tyr63 and Ile66 for most of the simulations. It can be suggested that the peptide remained “trapped” in this conformation due to the continuous interaction with the lipids.
A similar trend was observed in the simulations starting from a helical conformation (Figure 7.9). With increasing lipid concentration an unfolding of the helix and formation of a β-hairpin structure was observed. Interestingly, in the system 4lh, the helical content of the peptide was retained for a longer period. During this simulation, there was a continuous interaction between the lipids and the peptide (described in Section 7.4.1), suggesting that the lipids have a stabilising effect on the peptide structure. In contrast, at lipid concentration of 5:1 for the 5lh system almost immediate unfolding was observed, however the system was restrained by the lipids in a β-strand conformation. This observation indicated that there is a minimum number of lipids required to interact with the peptide in order to restrict the peptide’s conformation. Similarly to the 4lh system the restricted rate of unfolding in the peptide at higher lipid concentrations was further emphasised in the system 6lh. During the 300 ns simulation, the peptide remained in a helical conformation for ~90% of the total time (Figure 7.9, 6lh), having at least 4 lipids bound to it at any particular time (Section 7.4.1). The limited unfolding and conservation of the conformation of apoC-II(60-70) peptide for this system can be seen in the structural model shown in Figure 7.8. To investigate the statistical significance of this event, a second simulation of 300 ns was undertaken. In the second simulation at 6:1 concentration (6lh2), the peptide retained its helical content during the initial 150 ns, followed by the formation of a consistent turn structure for the rest of the simulation.

The importance of the lipid-peptide interactions was highlighted with the significant difference in secondary structure observed in the systems with 4:1 to 6:1 lipid:peptide ratio, compared to the consistent trend observed for system with 1:1 to 3:1 lipids per peptide.

Fewer occurrences of β-hairpin conformation were found in the presence of 3 or more lipids per peptide. This suggests that a possible role of the lipid molecules at high concentrations in inhibiting the aggregation process may be that it prevents the peptide from forming a β-hairpin structure and helps it maintain an extended conformation. Moreover, the “coating” effect which traps the peptide conformation at higher lipid concentrations, as seen in our simulations, can have an inhibitive effect on fibril formation. At these concentrations the peptide is less flexible and has reduced exposure for further self-association. Circular dichroism (CD) data also suggested that micellar lipids induce α-helical conformation in full-length apoC-II, which in turn suppresses the propensity to aggregate into amyloid fibrils [142]. Furthermore, the trapping effect of lipids on the conformation of apoC-II(60-70) dimers was previously illustrated by experimental and computational techniques [285].
The reduced conformational fluctuation with increasing lipid concentration was observed by analysing the cluster formation over time (figure shown in the Appendix of the thesis). Lower numbers of clusters were seen as the lipid concentration was increased, suggesting higher conformational stability. In particular, the lowest number of clusters formed for the 6lh system, where the peptide was trapped in a mostly helical conformation by the surrounding lipids.

Overall, regardless of the starting conformation of apoC-II(60-70), faster relaxation of the peptide structure was seen at low lipid concentration (1 and 2 lipids per peptide). Significant differences in structural transformations were noted at lipid:peptide ratios of 3:1 and above, suggesting much slower peptide mobility resulting from lipid binding. The degree of change in secondary structure of the peptide depends on the lipid concentration, i.e. increased flexibility and loss of intra-molecular interaction was observed as the lipid concentration was raised. However, at concentrations of 4 lipids and higher per peptide, lipids were able to coat the peptide and “trap” it in a particular conformation, restricting its dynamics and, possibly reducing the self-association ability of the peptide.

**Figure 7.8** Structural models of several lipid-peptide complexes. The surface of the peptide is coloured yellow, whereas the surface of the lipids is coloured red.
CHAPTER 7. Lipid concentration effects on the amyloidogenic apoC-II(60-70) peptide

Figure 7.9 Secondary structure evolution of apoC-II(60-70) at various lipid concentrations. Each simulation is labelled as described in the Computational Details section of this chapter.
7.4.3 Umbrella sampling: Free energy profiles

To examine the effect of lipid concentration on the unfolding of the apoC-II(60-70) peptide, quantitative analysis was performed by implementing the umbrella sampling technique. For the umbrella sampling simulations the coil structure of apoC-II(60-70) was selected as a starting conformation in order to investigate the free energies of the transitions from β-hairpin to an extended β-strand in the presence of lipids. In Chapter 6 we showed that the equilibrium structure of apoC-II(60-70) was a β-hairpin conformation, suggesting that this structure may be the intermediate state of the peptide monomer prior to self-association into fibrils.

The potential of mean force (PMF) profiles as a function of centre of mass (COM) separation between the Cα atoms of the N- and C-termini at various lipid concentrations were constructed and are presented in Figure 7.10. Due to the close proximity of the free energies, magnified representation of the PMF profiles at the centre of mass separation regions of 0.35 nm - 0.6 nm and 1.2 nm - 1.6 nm were taken and are shown as in inset in Figure 7.10.

![Figure 7.10](Image)

**Figure 7.10** Potential of mean force (PMF) profiles as a function of centre of mass separation (COM) of the N- and C-termini of apoC-II(60-70) peptide at various lipid concentrations. Magnified representations of selected regions are shown as an inset of the figure. Each PMF profile is coloured and labelled as described in the *Computational Details* section of this chapter.
The PMF profiles exhibit a linear decrease in energy as the N- and C- termini are being separated in the presence of lipids. However, close inspection of the difference in energy shows that higher energy is required to separate the termini with increasing lipid concentration. In particular, at 2 lipids per peptide and higher, a systematic difference in energy is shown in the inset on the right in Figure 7.10. Interestingly, significantly lower energy was required to obtain the initial termini separation from 0.3 nm to 0.7 nm for system 1lc (green, Figure 7.10) compared to the ambient system in lipid-free environment (red, Figure 7.10). This result suggests that the lipid destabilised the intra-molecular hydrophobic interactions within the core of the peptide, thus allowing it to extend easily. As was demonstrated in Chapter 6, under ambient conditions the wild-type apoC-II(60-70) peptide exhibits a compact β-hairpin structures, which are stabilised by strong hydrophobic interactions between the hairpin strands Ser61-Tyr63 (strand 1) and Phe67-Asp69 (strand 2). The energies obtained for these two systems are in agreement with the findings from our analysis of the lipid-peptide interactions where a higher number of interactions was made with the hydrophilic residues by the lipid in the 1lc system, as described in Section 7.4.1. We propose that the single lipid destabilises the hydrophobic interaction within the peptide, which enables the peptide strands to be easily separated. Thus, a lower energy was observed for the initial separation of the termini of the apoC-II(60-70) peptide in a single lipid (1:1) environment compared to the lipid-free environment, as expected, and this correlates well with our MD simulations.

Additionally, the higher energy required for inter-termini separation of apoC-II(60-70) at increased lipid concentrations (3 or more lipids) suggests that the binding of lipids restricts the peptide’s propensity for conformational change. These results give further support to our observations discussed in Sections 7.4.2, that lipids restrain apoC-II(60-70) dynamics and can trap the peptide in an arbitrary conformation.

7.4.4 Umbrella sampling: Lipid-peptide interactions

Similarly to the analysis performed on the data collected using molecular dynamics, we were interested in determining the type of lipid-peptide interactions that occurred in the biased simulations obtained using umbrella sampling. The solvent accessible surface (SAS) area per residue was calculated for each system to identify favourable interactions sites for the lipids.
Performing the same analysis on the trajectories accumulated using umbrella sampling allowed us to determine if the type of conformations the peptide exhibits will have an effect on the preferred interaction sites of the lipids. A probe of 1.4 Å radius and Lennard-Jones hard-shell radii for each atom were used to define the SASA. Figure 7.11 illustrates the average SASA and corresponding error per residue for each system investigated using the umbrella sampling technique.

The SASA analysis performed on the trajectories obtained using umbrella sampling exhibited no significant difference to those obtained in equilibrium using classical molecular dynamics. Overall, the results showed a decrease in SASA as the lipid concentration was increased. Similarly to the results obtained from the molecular dynamics simulations in Section 7.4.1, the preferential interaction sites of the lipids were the aromatic residues (Tyr63 and Phe67) and their neighbours, indicating that the apoC-II(60-70) peptide conformation does not affect the preferred location for lipid interactions.

Furthermore, the radial distribution functions (RDF) of water with respect to the peptide in the presence and absence of lipid were also calculated for the data obtained by umbrella sampling. The analysis of the solvation of apoC-II(60-70) was performed on the umbrella sampling simulation windows in the region of lowest free energy (Figure 7.10). The conformations in these simulation windows exhibit an extended β-strand structure with inter-termini COM distance of ~1.9 nm. The RDF profiles of water in the lipid-free and various lipid concentrations are shown in Figure 7.12.

The RDF profiles of the water with respect to the apoC-II(60-70) peptide in region of conformations with lowest free energy (extended β-strand) exhibited similar results to the RDFs obtained from the MD simulations (Section 7.4.1). The higher peak in g(r) was seen for the peptide in a lipid-free environment at ~1.82 Å, which is expected as there are no lipids to intervene in the hydrogen bonding of water to the peptide. At longer separations, the lowest density of water was found for the system with the highest lipid concentration of 6:1 lipids per peptide.
Figure 7.11 Average solvent accessible surface area (SASA) per residue. The calculations were performed on the data collected using umbrella sampling. Each plot is labelled as described in the Computational Details section of this chapter.
CHAPTER 7. Lipid concentration effects on the amyloidogenic apoC-II(60-70) peptide

Figure 7.12 Water radial distribution functions (RDF) with respect to apoC-II(60-70) peptide at various lipid concentrations calculated from the simulations performed using umbrella sampling at the lowest free energy region. Magnified representation of the region of the first peak is shown as inset. The RDF profiles are coloured as described in the Computational Details of this chapter.

7.5 Conclusions

In this chapter we investigated the effect of short-chained D5PC phospholipid at various concentrations in aqueous solution, on the structure and dynamics of apoC-II(60-70) peptide. Classical molecular dynamics simulations were performed to examine the type of lipid-peptide interactions, and their effect on the conformation of the peptide. The umbrella sampling technique was also applied to obtain a more quantitative analysis of the conformational free energies with respect to lipid concentration. Our results suggested that different lipid concentrations can have significant impact on the structure and stability of apoC-II(60-70) monomers.
The analysis of the lipid-peptide interactions showed that the highest number of interactions are occurring between the lipids and the hydrophobic residues (Tyr63, Gly65, Ile66, Phe67) of apoC-II(60-70). The head component of the lipid was found to interact with all residues, however more favourably with the hydrophilic regions, while the lipid tails prefer to interact with the hydrophobic region. The observed presence of lipid heads in the region of hydrophobic residues could suggest a micellar-like formation where the tails are buried in the peptide while the heads are exposed to the solvent.

The average solvent accessible surface area of the peptide showed that there is a preference by the lipids to interact with the aromatic residues (Tyr and Phe) and their neighbouring residues. The interaction with the hydrophobic residues was shown to reduce the dynamic fluctuations of the peptide, particularly at high lipid to peptide ratios of 4:1 to 6:1. These preferred sites for interactions were the same in all the simulations performed in this work, suggesting that the specific conformation of the peptide does not influence the favourable binding sites of the lipid. The shielding of the peptide’s aromatic rings by lipids can contribute to a fibril inhibiting mechanism whereby the peptides are prevented from self-association.

Significant change in the secondary structure of apoC-II (60-70) peptide was observed with increasing lipid concentration. At lower concentrations (1-3 lipids per peptide) contacts with the lipids reduced the intra-molecular interactions within the peptide, which allowed the peptide to easily unfold and adopt extended β-strand conformations. In contrast, higher lipid concentration (4-6 lipids per peptide) had a restraining effect on the peptide’s flexibility by trapping it in a particular conformation. This behaviour can also be suggested as inhibiting fibril formation, because of the lipid induced peptide’s inability to adopt conformations favourable for further self-association. However, this behaviour was found to be dependent on the number of lipids bound to the peptide at a given time.

Using the umbrella sampling method we were able to determine the free energy of the peptide extension, starting from a coil conformation, as a function of the lipid concentration. We found that higher energy is required to extend the peptide with increasing lipid concentration, which suggests that lipids have a restricting effect on the peptides dynamics. These results are consistent with our unbiased molecular dynamics simulations.

Finally, the structuring of water in the proximity of the peptide was analysed using radial distribution functions. The results obtained showed high density of water around the
peptide at low lipid concentration (1 to 3 lipids per peptide), while lower water presence was found in the system with high lipid concentration (4 to 6 lipids per peptide). The reduced solvation of the apoC-II(60-70) peptide at the higher lipid concentrations suggests that the lipids are coating the peptide while they expel the water from its surroundings, thus restricting the conformational transformations of the peptide.
Chapter 8

8. Oligomers of apoC-II(60-70): stability & effect of mutation

8.1 Overview

The oligomeric intermediates are believed to be the cytotoxic species in diseases associated with amyloid deposits. Therefore the mechanisms of formation and stability of these oligomeric species are crucial to understand.

Continuing from our studies of the monomeric apoC-II(60-70) peptide, MD simulations of apoC-II(60-70) oligomeric assemblies of various sizes and arrangements were performed. Specifically, we investigated the structural stability of dimeric, trimeric and tetrameric single $\beta$-sheet formations in parallel and anti-parallel strand orientations. The effect of different terminal states, e.g. charged, NH$_3$+...COO-, and neutral, NH$_2$...COOH was examined. The effects of mutations, such as oxidation of Met60, Met60Val and Met60Gln, on the conformation of the most stable oligomer were also studied. To acquire insight into the initial stages of self-association umbrella sampling simulations were performed to determine the dimerisation free energies of the wild-type and mutated apoC-II(60-70) dimers, with charged and neutral termini.

It should be noted that the simulations described in this chapter will be extended in the future and results presented here should be treated as preliminary.
8.2 Introduction

Recent findings have suggested that the precursors to amyloid fibrils, such as low-molecular-weight oligomers, and/or structured protofibrils, are the cytotoxic species in the amyloid diseases [145]. The elucidation of their mechanisms of formation and structural stability is important, as it will aid the development of therapeutics that could prevent their formation or enable their segregation. Experimentally, it is well established that fibril growth is initiated by a minimal pre-formed soluble, self-assembled seed complex or nucleus that must retain its structural organisation in the solvent in order to serve as a template for the fibril growth process. Thus, rather than sampling many micro-states of aggregation of small molecular assemblies, recent computational studies have been exploring possible final arrangements of the oligomeric complexes and evaluating their relative stabilities.

Using this approach, Nussinov et al. were able to interpret some experimental results and suggested an elongation mechanism for the protofilament of the Syrian hamster prion protein, ShPrP(113-120) [124]. They were also able to rationalise the observed structural preferences of Aβ peptide segments [112], and predicted the micro-structural differences of the islet amyloid polypeptide segments 22-27 and 22-29 [357, 358]. In all the peptides they investigated, the organisation of the smallest ordered structure always implied lateral association [357].

In a recent computational study performed by Jiang et al. the effect of lipids on the formation of fibrils of hIAPP(20-29) peptides was investigated [347]. The stability of several oligomeric formations was of particular interest. They found that only three out of the ten residues in each strand form hydrogen bonds which are competent to stabilise a tetrameric sheet, while the lipid served as a catalyst for the stability of the β-sheets. Additional computational studies that have investigated oligomeric filaments in various environments are summarised in Chapter 1.

In this chapter, the conformational stability of pre-formed oligomeric species of apoC-II(60-70) of various sizes and arrangements were investigated. The dimer, trimer and tetramer were studied, in a parallel and anti-parallel β-strand orientation. The effect of terminal charges on the stability of each oligomer was investigated. Once the oligomer with the higher stability was determined, the effect of oxidation of Met60 and single point mutation to Met60Val and Met60Gln on the structural resilience was studied. Furthermore, umbrella sampling
simulations were performed to determine the free energies of the initial stages of aggregation, specifically on the dimer formation of apoC-II(60-70). The effect of mutation (oxiMet60, Met60Val and Met60Gln) and terminal charge on the dimerisation free energy was also investigated. Some preliminary results of these calculations are presented in this chapter.

8.3 Computational Details

The structural stability of several oligomeric filaments of apoC-II(60-70) was investigated using classical molecular dynamics simulations. Similar to all the previous investigations, the Gromacs suite of programs [209] was employed to perform classical molecular dynamics calculations in conjunction with the Gromos 43A1 force field. Each system was energy minimised to remove steric clashes using the steepest descent algorithm with convergence criterion of 20 kJ/mol/nm. To allow the solvent to relax around the peptide oligomer, equilibration of 1 ns was performed with the oligomer restrained. The simulations were performed in the isothermal-isobaric (NPT) ensemble. Constant temperature at 300 K was achieved by coupling the system to a Berendsen thermostat, and constant pressure of 1 bar was maintained using the Berendsen barostat [188]. Atom based cutoff of 10 Å was used to treat the nonbonded van der Waals interactions. The Particle Mesh Ewald (PME) summation method [185] was applied to treat the long-range electrostatic interactions. The SHAKE algorithm was used to constrain all bond lengths to their equilibrium value, enabling a time step of 2 fs to be used for all calculations.

The oligomers examined were dimers, trimers and tetramers, in a single β-sheet arrangement. Two different strand orientations were examined, parallel (P) and anti-parallel (AP). The effect of termini charges on the stability of the formed oligomers was examined by evolving each system with charged (C), NH3+…COO-, and neutral (N), NH2…COOH termini ends. A schematic of the β-strand arrangement and size of the oligomers under investigation is shown in Figure 8.1.
CHAPTER 8. Oligomers of apoC-II(60-70): stability & effect of mutation

The oligomeric species were built with apoC-II(60-70) peptides in an extended conformation (β-strand). The strands were ~0.77 nm apart from one another, minimising initial inter-strand interactions. The initial β-strand distance separation was selected to reduce the need for extensive computational sampling, but still allow the system some freedom to adopt the favoured orientations. Each oligomeric system was enclosed in a periodic cubic box with explicitly represented SPC water molecules, corresponding to water density of ~1.0 g/cm$^3$. The apoC-II(60-70) peptide has an overall charge of -1, so a single counterion Na$^+$ was added for every strand present in the simulation box. To help in the interpretation of the results, Table 8.1 shows a summary of system names, box sizes, the number of water molecules and counterions present in each simulation box, and the total simulation time accumulated for each system.

The first part of our investigation involved the determination of the most stable oligomer formed. Once the stable oligomer was identified, single point residue substitutions including the oxiMet, Met60Val and Met60Gln mutations (as implemented in Chapter 6), were performed to examine their effects on the stability of this oligomer.

To obtain a more quantitative illustration of the self-association mechanisms of dimeric apoC-II(60-70) and effects of mutations, umbrella sampling simulations were performed. The umbrella sampling [200] methodology was applied to calculate the free energy ($\Delta G_{\text{dimer}}$) of formation of a dimer from two monomers as a function of their centre-of-
mass (COM) separation. The separation between 0 and 2.25 nm was maintained in each window by Hookean functions with force constant of 20 kcal/mol/Å², separated from adjacent windows by 0.05 nm. Forty five windows were simulated for 3 ns each. The Weighted Histogram Analysis Method (WHAM) [349] was applied to obtain the unbiased free energy profiles (potentials of mean force, PMF) of dimerisation of apoC-II(60-70).

<table>
<thead>
<tr>
<th>System Name</th>
<th>Sheet, #Strands, Arrangement</th>
<th>Simulation Box Size (Å)</th>
<th>Number of Waters</th>
<th>Number of Counterions</th>
<th>Simulation Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native structure oligomers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-2-PC</td>
<td>61 × 61 × 61</td>
<td>7480</td>
<td>2 Na+</td>
<td>100 ns</td>
<td></td>
</tr>
<tr>
<td>1-2-PN</td>
<td>61 × 61 × 61</td>
<td>7487</td>
<td>2 Na+</td>
<td>100 ns</td>
<td></td>
</tr>
<tr>
<td>1-2-APC</td>
<td>61 × 61 × 61</td>
<td>7479</td>
<td>2 Na+</td>
<td>100 ns</td>
<td></td>
</tr>
<tr>
<td>1-2-APN</td>
<td>61 × 61 × 61</td>
<td>7483</td>
<td>2 Na+</td>
<td>100 ns</td>
<td></td>
</tr>
<tr>
<td>1-3-PC</td>
<td>63 × 63 × 63</td>
<td>8074</td>
<td>3 Na+</td>
<td>100 ns</td>
<td></td>
</tr>
<tr>
<td>1-3-PN</td>
<td>63 × 63 × 63</td>
<td>8076</td>
<td>3 Na+</td>
<td>100 ns</td>
<td></td>
</tr>
<tr>
<td>1-3-APC</td>
<td>63 × 63 × 63</td>
<td>8069</td>
<td>3 Na+</td>
<td>100 ns</td>
<td></td>
</tr>
<tr>
<td>1-3-APN</td>
<td>63 × 63 × 63</td>
<td>8074</td>
<td>3 Na+</td>
<td>100 ns</td>
<td></td>
</tr>
<tr>
<td>1-4-PC</td>
<td>65 × 65 × 65</td>
<td>8859</td>
<td>4 Na+</td>
<td>100 ns</td>
<td></td>
</tr>
<tr>
<td>1-4-PN</td>
<td>65 × 65 × 65</td>
<td>8850</td>
<td>4 Na+</td>
<td>100 ns</td>
<td></td>
</tr>
<tr>
<td>1-4-APC</td>
<td>65 × 65 × 65</td>
<td>8846</td>
<td>4 Na+</td>
<td>100 ns</td>
<td></td>
</tr>
<tr>
<td>1-4-APN</td>
<td>65 × 65 × 65</td>
<td>8847</td>
<td>4 Na+</td>
<td>100 ns</td>
<td></td>
</tr>
<tr>
<td>Mutated Oligomers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-4-APC-M60Q</td>
<td>65 × 65 × 65</td>
<td>8895</td>
<td>4 Na+</td>
<td>100 ns</td>
<td></td>
</tr>
<tr>
<td>1-4-APC-M60V</td>
<td>65 × 65 × 65</td>
<td>8899</td>
<td>4 Na+</td>
<td>100 ns</td>
<td></td>
</tr>
<tr>
<td>1-4-APC-oxiMet</td>
<td>65 × 65 × 65</td>
<td>8894</td>
<td>4 Na+</td>
<td>100 ns</td>
<td></td>
</tr>
</tbody>
</table>

**Table 8.1** System set-up listing the names given for each simulation, together with the simulation box dimensions, number of water molecules, counterions and total simulation time.
8.4 Results and Discussions

8.4.1 Root mean square deviation and clustering analysis of wild-type apoC-II(60-70) oligomers

The stability of each oligomer was monitored by calculating the root mean square deviation (RMSD) of the backbone with respect to the starting conformation over the entire trajectory. In order to determine the most populated structure of every oligomer, clustering analysis was performed using every second frame of the trajectory (12500 frames). The RMSD plot together with the most populated structures and their population for the dimer, trimer and tetramer of apoC-II(60-70) are presented in Figures 8.2, 8.4 and 8.5, respectively.

Close inspection of the RMSD plot in Figure 8.2 showed a progressive increase in RMSD in the first 20 ns of the simulations, suggesting a significant change in conformation. The highest RMSD values were exhibited by the dimers in parallel arrangement, 1-2-PC and 1-2-PN having an average RMSD of 1.0 nm and 0.95 nm, respectively. The anti-parallel dimers exhibited lower RMSD values, with 1-2-APC and 1-2-APN having an average RMSD of 0.69 nm and 0.76 nm, respectively. The significant change in the parallel structure compared to anti-parallel dimers was demonstrated by the lower population of structures in the most stable parallel clusters, 47.9% and 43.6%, versus 93.7% and 67.4% for the anti-parallel clusters. These results indicated that the parallel arrangement of apoC-II(60-70) dimers is less favourable than the anti-parallel, regardless of the termini’s charge. Both arrangements did not retain the extended β-sheet starting structure, but rather formed a β-hairpin-like conformation, where higher number of intermolecular interactions were made by the peptide in the anti-parallel arrangement, as discussed below.

The interactions between the residues were examined by constructing a residue contact map for the backbone and side-chain atoms individually, in order to determine which component contributes most to the structures formed. The contact maps shown in Figure 8.3 present the backbone and side-chain distances between individual residues averaged over the structures sampled in each apoC-II(60-70) dimer trajectory.
Figure 8.2 Root mean square deviation (RMSD) plots of the trajectories of dimeric apoC-II(60-70) oligomers. The structures of the most populated clusters of every system and their population are also presented.
Figure 8.3 Contact maps representing the average contact distances between the backbone and side-chain atoms between each residue of each dimeric system of apoC-II(60-70). The residues of monomer 1 are shown on the horizontal axis, whereas the residues of monomer 2 are shown on the vertical axis.
The contact plots showed larger separations between the two monomers of the 1-2-PC system. Clustering of close contacts can be seen near the central region of the two peptides, between residues Gly65 to Phe67, while the termini are not interacting strongly, exhibiting separation of more than 0.5 nm. The 1-2-PN system exhibited similar backbone contact distances to the 1-2-APC and 1-2-APN system, where close backbone contacts between the residues can be seen in a diagonal arrangement, suggesting anti-parallel configuration. This result suggests that the parallel dimer has rearranged to form an anti-parallel-type orientation, even though it is not in an extended β-sheet state. Interestingly, there are more close contacts seen between the backbone atoms, compared to the side-chain groups, which could suggest that in the apoC-II(60-70) dimers the backbone-backbone contacts are responsible for the β-sheet formation, while the side-chain-side-chain contacts contribute to the fibril stability. Close side-chain contacts have formed between the aromatic residues, in particular in the anti-parallel system of 1-2-APC. More recently, the role of aromatic residues in the fibril forming process and subsequent fibril stability has been shown to be significant [130].

The simulations of the apoC-II(60-70) trimers exhibited similar results to the dimer systems. The least stable structures were those of the parallel oligomers, as demonstrated by the highest RMSD values (Figure 8.4). The average RMSD of the 1-3-PC and 1-3-PN was 1.12 nm and 0.96 nm, respectively. In the simulation of the 1-3-PC system, a disruption of the parallel β-strand alignment was observed by one of the external strands. Rearrangement between the strands was seen where the outer strand separated from the dimer and rotated to reattach back in an anti-parallel orientation. The structure was retained for the rest of the simulation, and can be seen in the most stable cluster of system 1-3-PC represented in red of Figure 8.4. This result gives further insight into the possible orientation preference for fibril formation in the apoC-II(60-70) peptide. Similar to our results, the transition from parallel to anti-parallel sheet has also been captured in the replica exchange studies of Jiang et al. [347]. They demonstrated that for the transition from parallel to anti-parallel orientation a complete detachment of the strand is not required. Their results showed that the transition involved only internal reorganisation where one single hydrogen bond in the parallel pattern remained while the whole strand rotated by 180° around the hydrogen bond. Afterwards newly generated anti-parallel hydrogen bonds formed near the place of the original hydrogen bond.
Figure 8.4 Root mean square deviation (RMSD) plots of the trajectories of trimeric apoC-II(60-70) oligomers. The structures of the most populated clusters of every system and their population are also presented.
The anti-parallel trimers were more stable compared to the parallel, having an average RMSD of 0.87 nm and 0.69 nm for the 1-3-APC and 1-3-APN systems, respectively. A stable dimer formation can be seen in both anti-parallel trimers, having an extended β-sheet structure. The third strand of every system was found to be interacting less with the stable dimer. Further investigation in the type of interactions and consequent structural changes of the trimeric oligomers will be conducted.

The results obtained from the apoC-II(60-70) tetramer simulations exhibited in general higher structural stability in comparison to the dimer and trimer systems (Figure 8.5). The parallel simulations, 1-4-PC and 1-4-PN exhibited progressive instability, where the external strands were slowly starting to dissociate from the dimer. Similarly to the 1-3-PC system (Figure 8.4), transition from parallel to anti-parallel orientation was seen to occur by one of the outer strands of the parallel tetramer. Both of these systems have charged termini, which could possibly contribute to the transition, since initially the strands are repelled from one another, then undergo some conformational changes, and finally are attracted via the oppositely charged ends, to form an anti-parallel structure. However, regardless of the charge state of the termini, the parallel arrangement of the apoC-II(60-70) tetramer is less stable than the anti-parallel orientation of the β-strands. The average RMSD of the parallel, 1-4-PC and 1-4-PN simulations, was 0.85 nm and 0.65 nm, respectively, whereas the anti-parallel, 1-4-APC and 1-4-APN simulations had an average RMSD of 0.35 nm and 0.48 nm, respectively.

The anti-parallel tetramers exhibited extended β-sheet conformation with high stability, represented by the high population of the most stable cluster structure shown in Figure 8.5. The average centre-of-mass separation over the entire trajectory between the four strands (1-2, 2-3, 3-4) for the 1-4-APC system was 0.47 nm, 0.47 nm and 0.48 nm, respectively, while for the 1-4-APN system the average distances were, 0.35 nm, 0.52 nm and 0.31 nm, respectively. These inter-strand distances are in correlation with the experimentally determined values of ~0.47 nm [110] found in most amyloid fibril structures. Additional analysis, such as hydrogen bond patterns, backbone and side-chain contact maps and orientation will be conducted on all simulations to better understand the specific factors that contribute to these oligomer formations and their stability.
Figure 8.5 Root mean square deviation (RMSD) plots of the trajectories of tetrameric apoC-II(60-70) oligomers. The structures of the most populated clusters of every system and their population are also presented.
From the investigations conducted thus far, our simulations showed that an increase in the number of strands improved the stability of the oligomers dramatically, regardless of the strands’ orientation. However, the anti-parallel arrangement between the strands was more favourable, as indicated by the lower RMSD compared to the parallel oligomers, and the observed change in strand orientation of some parallel systems, such as 1-2-PN and 1-3-PC.

8.4.2 Root mean square deviations of mutated apoC-II(60-70) oligomers

In the previous section, we identified that the tetrameric apoC-II(60-70) in anti-parallel arrangement, system 1-4-APC, was the most stable oligomeric species. The termini charge showed no significant effect on the stability of this oligomer. We proceeded to investigate the effects of mutation on the structural stability of the most stable oligomer (1-4-APC) by performing single residue substitutions at the location of Met60. The same mutations that were implemented to study the mutation effect on monomeric apoC-II(60-70) in Chapter 6, were also applied for this study. The effect of oxidised Met60 (oxi-Met), and mutations to hydrophobic Val (Met60Val) and hydrophilic Gln (Met60Gln) was investigated. For clarity each system was given a name based on the mutation performed, as listed in Table 8.1. The conformational stability of the mutated oligomers was monitored by calculating the RMSD with respect to their starting conformation, over the entire trajectory. For comparison, the RMSD graph of the wild-type oligomer was reproduced and is presented together with the mutated simulation results in Figure 8.6. Several structures representing typical conformations observed during the simulations of the mutated systems are shown in Figure 8.7.

During the 100 ns of simulation of each system the oligomers retained their β-sheet structure with some differences in their conformation. The strong stability of the β-sheet structure can be attributed to the presence of strong hydrophobic core between the aromatic residues, Tyr63, Gly65, Ile66 and Phe67. The hydrophobic core surface is illustrated on one of the typical structures observed for the 1-4-APC-oxiMet oligomer in Figure 8.7a. Furthermore, the presence of two aromatic residues, Tyr63 and Phe67 enhanced the durability of the β-sheet by π-stack ring formations throughout the entire trajectory of every system. The aromatic ring alignment is shown on the typical structures observed for the 1-4-APC-oxiMet and 1-4-AP-M60V system, depicted in Figure 8.7a, b. This confirms the role of the aromatic
rings in fibril formation to be significant, as discussed in Chapter 6, specifically in relation to their presence in fibril forming regions and their contribution to fibril stability [130].

Close inspection of Figure 8.6 showed similar trend of RMSD fluctuations between the wild-type 1-4-APC and 1-4-APC-M60V systems. There was no considerable change in the peptide arrangement or the overall structure of the oligomers in these two systems. The conformity of their typical structures can be seen in Figure 8.5 and 8.7b. In contrast, a more significant change in structure was observed at 15-20 ns of simulation of the 1-4-APC-oxiMet and 1-4-APC-M60Q systems, as indicated by the steep increase in RMSD during this time. Close inspection of the simulated trajectories at these times showed a twisting motion of β-sheet structure. A typical “twisted” conformation observed for the 1-4-APC-M60Q system is depicted in Figure 8.7c. The illustration shows central β-sheet formation with fraying in the N- and C-termini ends. The twist and fraying at the ends can be attributed to the increased hydrophilicity of the residues at the N-terminus, however additional analysis to clarify the formation of this structure will be conducted.

Twisting in β-sheet structures has been observed for other peptides ([359] and references therein). From a series of MD simulations performed under different conditions on the cross-β filament of the GNNQQNY peptide, Pariole et al. found that in the absence of crystal packing interactions there is no free energy barrier for the twisting of the filament. Additionally, entropic contributions, in particular those associated with an increase in backbone dynamics upon twisting, stabilise the twisted form.

During the 100 ns of simulations of each mutated system the oligomers did not dissociate, in good correlation with the experimental observations of mutated apoC-II peptides, which have showed a concentration dependent behaviour for fibril formation. It was observed that at high peptide concentrations the mutated apoC-II protein and derived peptide (same mutations as applied in this study), can form fibrils (unpublished/personal communication), even for the oxidised peptide which have been shown to inhibit fibril formation at lower concentrations [285]. Our simulations indicate that once the fibril forming nucleus has been formed, the oligomers are stable, at least in the timeframe of the current simulations (100 ns).
Figure 8.6 Root mean square deviation (RMSD) plots of wild-type and mutated, four stranded anti-parallel oligomer of apoC-II(60-70).
Figure 8.7 Structures representing the typical conformations observed during the simulations of mutated tetrameric apoC-II(60-70) oligomers; a) 1-4-APC-oxiMet (red), b) 1-4-APC-M60V (green) and c) 1-4-APC-M60Q (blue). In some structures the aromatic side-chains are also shown, as well as the surface of the hydrophobic residues is represented as a green mesh.
8.4.3 Free energy profiles of wild-type and mutated apoC-II(60-70) dimers

To obtain a more quantitative description of the dimerisation process of the apoC-II(60-70) peptide, and to understand the effect of mutations on the initial stages of self-association, umbrella sampling simulations were performed. The free energy ($\Delta G$) profiles of dimerisation as a function of centre-of-mass (COM) separation to 2.25 nm of the two monomers were determined. The effect of charged and neutral termini was also investigated, and the free energy profiles are shown in Figure 8.8.

Preliminary interpretation of the data presented showed overall higher free energies required to separate the systems with charged termini (Figure 8.8a) compared to the systems with neutral termini (Figure 8.8b). This result can be expected, because of the attractive arrangement of the opposite charges that form in an in-register anti-parallel peptide dimers. Thus, higher energy is required to separate the anti-parallel dimers with charged termini, compared to the dimers with neutral termini, regardless of the mutation at the N-terminus. However, some differences between the dimerisation energies of the mutated systems were observed.

The dimerisation free energies for the apoC-II(60-70) dimers with charged termini (Figure 8.8a), exhibited distinct differences with respect to their mutations. The lowest free energy of separation was seen for the Met60Gln dimer, followed by the wild-type, Met60Val and highest energy was the apoC-II(60-70) dimer with oxidised Met60. The change between the dimerisation free energies of the mutated dimers with neutral termini was different compared to the dimers with charged termini. The lowest separation free energies were observed for the wild-type and oxidised dimers, while higher free energies were seen for the Met60Val and Met60Gln mutated dimers.

Inspection of the trajectories of each system showed continuous side-chain and termini interactions in the initial steps of separation of the dimer, thus additional analysis will have to be conducted to elucidate the distinct differences in free energies for each system. Residue contacts, hydrogen bond formation, aromatic ring orientations and termini contacts will be some of the calculations that will performed.
Figure 8.8 Free energy profiles as a function of the center-of-mass (COM) separation of the wild-type and mutated (oxiMet60, Met60Val and Met60Gln) apoC-II(60-70) dimers with a) Charged and b) Neutral termini.
CHAPTER 8. Oligomers of apoC-II(60-70): stability & effect of mutation

8.5 Conclusions

It is generally believed that the amyloid fibril formation follows a nucleated growth mechanism, where fibril growth occurs very fast once the critical nucleus seed has been formed. In this chapter the structural stabilities and dynamics of the apoC-II(60-70) dimer, trimer, and tetramer in parallel and anti-parallel arrangements using atomistic MD simulations were studied. The effects of termini charge and mutations were also investigated. From the preliminary analysis of the simulations conducted the following conclusions were made.

The preliminary results showed that the critical nucleus size may not be very large, since we observed that the tetrameric apoC-II(60-70) oligomer in anti-parallel configuration is very stable within the 100 ns of simulation. The termini charge had no effect on the stability and structure of the tetramer. The relatively small size of the nucleus seed of apoC-II(60-70) can explain the experimental observation of its rapid formation of amyloid fibrils [285].

The central hydrophobic core of apoC-II(60-70), between residues Tyr63 and Phe67, contributed to the stability of the tetramer. This finding was further emphasised by the stable β-sheet formation observed between these residues of each strand in the mutated apoC-II(60-70) tetramers. Oxidation and mutation of Met60 to Val and Gln did not appear to affect the stability of anti-parallel β-sheet of apoC-II(60-70). The π-stack alignment of the aromatic residues Tyr63 and Phe67 was consistent between the stable oligomers, and can be suggested to contribute to the stability of the pre-formed fibrils.

It should be stressed the results presented in this chapter are preliminary and are included to indicate the further directions of the investigation of this thesis. Additional quantitative analysis will have to be conducted to determine the type of interactions that contribute to the stabilities of the larger oligomers. Furthermore, detailed investigation of the residues (side-chain, backbone) and termini contacts of the dimer conformations accumulated from the umbrella sampling simulations will be performed to elucidate the differences in free energies observed between the dimers with charged and neutral termini.
Chapter 9

9. Future work

9.1 Computational techniques and force field development

With the continuous increase in computer power and technology, simulations of large systems at longer time-scales are becoming more feasible. However, the underlying factor in the validity of such simulations is the accuracy of the potential energy function. As presented in Chapter 4 of this thesis, the choice of force field in a biomolecular study is an important one, as it will ultimately reflect in the quality of the results. Therefore, continuous research into refinement and parameterisation of the force fields today is required.

The complexity of the free energy surface of a protein has caused many difficulties for theoreticians due to the inability to sufficiently sample the conformational space using classical molecular dynamics. The need for the development of a more sophisticated techniques capable of crossing large free energy barriers has become increasingly more evident. Recent examples include Bias-Exchange Metadynamics, Replica-Exchange Molecular Dynamics and Umbrella Sampling, to name a few. Specifically, the need for increased conformational sampling in the areas of protein folding (Chapter 5) and aggregation (Chapter 6, 7 and 8) is crucial, because important events, such as transition or intermediate states may not be accurately sampled using classical techniques. Continuous improvement and development of algorithms capable of sampling the entire conformational space of large protein complexes is necessary.
9.2 Theoretical investigation of fibril microcrystal structures

Recent advances in experimental techniques have enabled the determination of the microcrystal structures of 13 segments from fibril forming proteins using atomic-resolution crystallographic studies [110]. In this work, peptide segments containing 4 to 12 residues were discovered to form fibrils in isolation from the rest of the protein chain. These include segments from the Alzheimer’s amyloid-β and tau proteins, the PrP prion protein, insulin, islet amyloid polypeptide (IAPP), lysozyme, myoglobin, α-synuclein and β₂-microglobulin. It was discovered that the fundamental unit of these amyloid-like fibrils is a steric zipper, formed by two tightly interdigitated β-sheets. The orientation and arrangement of the β-strand and β-sheets within the fibrils varies between the different peptide segments.

The small size of these fibrils make them perfect peptide models for computational studies. The mechanisms of the fibril nucleation can be investigated in atomistic detail. In particular, the time evolved interactions that drive the nucleation process can be identified using classical molecular dynamics. Additionally, the environmental (pH, temperature, lipids) and mutational influences on the fibril forming propensities can also be investigated.

9.3 Lipids and apolipoprotein interactions in heart disease

Plasma levels of lipids and apolipoproteins are significant risk factors for coronary heart disease. Low levels of blood cholesterol, and particularly low density lipoprotein (LDL) cholesterol, correlate with a reduced cardiovascular risk [360, 361]. These important epidemiological findings underlie the widespread use of cholesterol lowering strategies as frontline measures in the prevention of heart disease [362]. Nevertheless, this disease continues to be a major killer of the human population. One of the great challenges currently posed for the development of better treatments for heart disease is to overcome the lack of understanding of the basic mechanisms that link blood lipids and apolipoproteins to the risk of heart disease. Plasma apolipoproteins transport blood lipids between organs and tissues. Lipids are also a common component of disease-related amyloid deposits and affect fibril formation by several lipid-binding and amyloidogenic proteins.
Our collaborators, Howlett et al. have used full-length apoC-II protein as a model to explore the effect of lipids on the self-assembly of apolipoprotein amyloid fibrils (summarised in Chapter 3). They hypothesised that specific apolipoprotein-lipid interactions control both the formation and morphology of apolipoprotein aggregates in atherosclerotic plaques. Characterisation of such interactions will help establish the mechanisms that link blood lipid levels with the risk of heart disease and suggest new ways to manage and reduce the incidence of heart disease.

The initial steps towards elucidation of the lipid-apolipoprotein interactions were made in our work presented in Chapter 7 of this thesis. The effects of lipids and their concentration on the conformation and dynamics of apolipoprotein derived peptide, apoC-II(60-70) were investigated. With this study we were able to improve our understanding in the lipid-peptide interactions that may accelerate or inhibit fibril formation of apoC-II(60-70).

An extension of this work will involve performing simulations of the full-length apoC-II protein in various lipid environments. A comprehensive theoretical approach can define the role of lipids in apolipoprotein mis-folding, self-assembly and amyloid fibril formation. Using theoretical and experimental techniques the following investigations can be suggested:

1) The effect of lipids on the conformation and self-assembly of apoC-II.

2) The structural specificity of lipids for their action on apolipoprotein fibril formation.

First, classical molecular dynamics can simulate the time evolution of the dynamics and interactions occurring between the protein and lipids at controlled conditions. Furthermore, docking and umbrella sampling simulations can be employed to obtain the free energies of lipid binding to the protein and identify the favourable interaction sites. Coarse-grained methods can be also applied to simulate the mesoscale dynamics of larger systems, composed of several proteins in the lipid rich solution. Lipid specificity can be explored by simulating various types of lipids.
9.4 Protein self-assembly on surfaces, interfaces and nanoparticles

A great challenge posed by the industrial uptake of nanomaterials is the lack of understanding of the interactions of these novel molecular systems with biological environments. Nanomaterials present a high surface area that can interact with proteins and promote a locally increased protein concentration leading to increased oligomer formation [363]. Surfaces presented by lipid bilayers [364] and liquid-air, liquid-solid, or liquid-liquid interfaces [365, 366] also exert specific and significant effects in promoting protein-protein interactions. A wide variety of new types of surfaces will emerge from the diversity of engineered nanoparticles that are designed for use in industrial applications, increasing the likelihood of human exposure.

As discussed in this thesis, protein aggregation is caused by the partial misfolding and self-association of proteins and can form insoluble amyloid fibrils, which appear in a number of common and debilitating diseases, including Alzheimer and Parkinson diseases. Surfaces presented by molecular chaperones, inhibit amyloid fibril formation [367], whereas lipid bilayers promote an alternate fibril morphology [142]. Similarly, surfaces presented by nanoparticles enhance the self-assembly of proteins into amyloid fibrils [363]. These observations demonstrate the potentially important role of surfaces, interfaces and nanoparticles in protein folding and the self-assembly of proteins into amyloid fibrils.

In this thesis we were able to show the effects of mutations and different environments on the structure and dynamics of the fibril forming apoC-II(60-70) peptide. Based on the findings made in this project and recent experimental discoveries, a three-year Australian Research Council (ARC) Discovery project has been granted to investigate the effects of various surfaces, interfaces and nanoparticles on protein self-assembly. The project will be in collaboration with Geoff Howlett and group, from the Bio21 Institute of University of Melbourne. Over the last 10 years, his team has extensively investigated the different processes that influence fibril formation in apoC-II (summarised in Chapter 3).

In this new project a comprehensive theoretical and experimental approach to study the role of surfaces, interfaces and nanoparticles in protein misfolding, self-assembly and amyloid fibril formation will be implemented. The specific aims of the project are to determine:
Future work

1) The mechanism of nucleation and self-assembly of misfolded peptides and proteins.
2) The effects of liquid-air and lipid interfaces on peptide and protein self-assembly.
3) The effects of surfaces and nanoparticles on protein-protein interactions and self-assembly.

The fundamental exploration of the role/s of “natural” biological interfaces, especially biomembranes, on the protein fibril-formation process is of great significance in the study of self-assembly of proteins. In particular, the exploration of the physiological effects of nanoparticle exposure and their connection with many conformation-related diseases. To the best of our knowledge, systematic investigations of the effects of surface chemistry, morphology and curvature on the molecular mechanisms of protein self-assembly and fibrillation have not been performed.
Appendix

Appendix 1. Radial distribution functions of the lipid head/s and tail/s with respect to hydrophobic and hydrophilic residues of the apoC-II(60-70) peptide at various lipid concentrations. Head-phobic (black), head-philic (red), tail-phobic (green) and tail-philic (blue).
Appendix 2. Histograms representing the contacts formed between the lipid head and tail with individual residues of apoC-II(60-70) at each lipid concentration. The percentages of time in contact with individual residue of the lipid head and tail are shown in the red/blue histograms, while the percentage of the number of contacts made with the residues by the lipid head and tail are shown in the grey/green histograms.

**Ilc**

![Ilc Time in contact (%)](image)

![Ilc Number of contacts (%)](image)

**Ilh**

![Ilh Time in contact (%)](image)

![Ilh Number of contacts (%)](image)
$3lh$

Time in contact (%)

Number of contacts (%)

$3lc$

Time in contact (%)

Number of contacts (%)
6lh2

![Bar charts showing the number of contacts and time in contact per day for different days of the week. The charts display the percentage of time spent in contact for head and tail positions.]
Appendix 3. Time evolution of cluster formation for each system. The percentage of population of the most stable cluster is written in red as an inset of each figure.
References


[94] R. Virchow, "Über eine im gehirn und rückenmark des menschen aufgefunden substanz mit der chemischen reaction der cellulose.", \textit{Archiv Fuer Pathologische Anatomie und Physiologie und Fuer Klinische Medizin}, \textbf{6}:135 (1854)

[95] H. Meckel, "Die speck-oder cholestrinkrankheit." \textit{Annalen des Charity Krankenhaus zu Berlin.}, \textbf{4}:264 (1853)


N.W. Isaacs and R.C. Agarwal, "Experience with fast fourier least-squares in refinement of crystal-structure of rhombohedral 2-zinc insulin at 1.5 Å resolution", *Acta Crystallographica Section A*, 34(SEP):782 (1978)


[244] I.S.G. Peking, "Insulin's crystals structure at resolution 2.5Å." *Peking Review*, 40:11 (1971)


