Understanding biopharmaceutical aggregation using minimalist models based on square-well potential

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy

in the Faculty of

Science & Engineering

2016

Hamza J. Magnier

School of Chemical Engineering And Analytical Science
# Contents

List of Figures ........................................................................... 7  
List of Tables ........................................................................... 9  
List of Abbreviations ................................................................ 11  
Physical Constants ..................................................................... 13  
List of Symbols ......................................................................... 15  
Abstract ....................................................................................... 17  
Declaration of Authorship ............................................................ 19  
Acknowledgements ...................................................................... 23  
1 Introduction .............................................................................. 1  
2 Background ............................................................................... 9  
2.1 Protein folding Problem .......................................................... 9  
2.1.1 Protein folding: a thermodynamic problem ............................. 9  
2.1.2 Protein folding: a kinetic problem ........................................ 10  
2.1.3 Protein folding: a hydrophobic core ..................................... 11  
2.2 Protein misfolding and aggregation ........................................... 12  
2.2.1 Aggregation in bioprocessing ............................................... 12  
2.2.2 Aggregation in human etiology ............................................ 14  
2.2.3 Aggregation types ............................................................ 14  
2.2.4 Physical-aggregation mechanisms ....................................... 15  
2.2.5 Kinetics of aggregation ....................................................... 18  
2.3 Simulation techniques for protein folding, misfolding, and aggregation ..................................................... 19  
2.3.1 The search problem .......................................................... 20  
2.3.2 Sampling ensembles ......................................................... 22  
2.3.3 Molecular Dynamics and Monte Carlo ................................. 24  
2.3.4 Other Sampling ensembles: replica-exchange technique ......... 25  
2.3.5 Protein Models: from all atoms to coarse-graining ............... 29  
2.3.5.1 High resolution models .................................................. 30  
2.3.5.2 Intermediate-resolution models ...................................... 31  
2.3.5.3 Low-resolution models .................................................. 33  
2.3.6 Interaction potential ........................................................ 38  
2.3.6.1 High resolution Interaction potential ................................. 38  
2.3.6.2 Intermediate- and low-resolution interaction potentials ...... 40  
2.3.7 Solvent types ................................................................. 45
3 Research Objectives
3.1 Effect of model parameter: finding a good model for studying protein behavior
3.2 Low-temperature structure: a new order-parameter to define the native ensemble
3.3 Investigating the correlation between the behavior of isolated peptides and aggregation propensity in multichain systems
3.4 Critical parameters calculation in square-well fluids

4 Approach
4.1 Model Description and Methodology
4.1.1 Model
4.1.2 Simulation Details
4.1.3 Interaction Potentials
4.2 Methodology
4.2.1 Temperature calculation
4.2.2 Temperature optimization
4.2.3 Data Analysis

5 Results and Discussion: effect of model parameters
5.1 Effect of number of beads ‘N’
5.1.1 Folding behavior of a 32 monomer chain
5.1.1.1 Model
5.1.1.2 A general overview
5.1.1.3 Folding transitions
5.1.1.4 Short Discussion
5.1.2 Folding behavior of a 64 monomer chain
5.1.2.1 Model
5.1.2.2 A general overview
5.1.2.3 Folding transitions and intermediates
5.1.2.4 A short discussion
5.1.3 Folding of longer chains
5.1.3.1 Model
5.1.3.2 A general overview
5.1.3.3 A short discussion
5.2 Effect of the overlap parameter ‘σ/l’
5.2.1 Model
5.2.2 64 monomer chain with λ equal to 1.5
5.2.3 64 monomer chain with λ equal to 1.1
5.2.4 Short discussion
5.3 Effect of range of interaction potential ‘λ’
5.3.1 General overview
5.3.2 Short discussion

6 Results and Discussion: low-temperature structures
6.1 Model
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.2</td>
<td>From a single native structure to a native ensemble: the rational of the approach</td>
<td>141</td>
</tr>
<tr>
<td>6.3</td>
<td>Overview</td>
<td>145</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Folding behavior</td>
<td>145</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Conformational fluctuations</td>
<td>150</td>
</tr>
<tr>
<td>6.4</td>
<td>Short discussion</td>
<td>154</td>
</tr>
<tr>
<td>7</td>
<td>Results and Discussion: multichain systems</td>
<td>157</td>
</tr>
<tr>
<td>7.1</td>
<td>Model</td>
<td>158</td>
</tr>
<tr>
<td>7.2</td>
<td>A single multichain system</td>
<td>159</td>
</tr>
<tr>
<td>7.3</td>
<td>Comparing multichain and isolated-chain systems</td>
<td>161</td>
</tr>
<tr>
<td>7.4</td>
<td>Short discussion</td>
<td>165</td>
</tr>
<tr>
<td>8</td>
<td>Results and Discussion: square-well fluids</td>
<td>167</td>
</tr>
<tr>
<td>8.1</td>
<td>Background</td>
<td>168</td>
</tr>
<tr>
<td>8.2</td>
<td>Model</td>
<td>171</td>
</tr>
<tr>
<td>8.3</td>
<td>Overview</td>
<td>172</td>
</tr>
<tr>
<td>8.3.1</td>
<td>Model fluids</td>
<td>172</td>
</tr>
<tr>
<td>8.3.2</td>
<td>Sticky-sphere model</td>
<td>175</td>
</tr>
<tr>
<td>8.4.2.1</td>
<td>Determination of critical parameters</td>
<td>175</td>
</tr>
<tr>
<td>8.4</td>
<td>Short discussion</td>
<td>178</td>
</tr>
<tr>
<td>9</td>
<td>Project conclusions</td>
<td>181</td>
</tr>
<tr>
<td>10</td>
<td>Recommendations for future work</td>
<td>185</td>
</tr>
<tr>
<td>A</td>
<td>Effect of interaction potential on a 64monomer chain with an overlap parameter of 1.3</td>
<td>187</td>
</tr>
<tr>
<td>B</td>
<td>Isolated chains low-temperature structure</td>
<td>191</td>
</tr>
<tr>
<td>C</td>
<td>Temperature lists</td>
<td>197</td>
</tr>
<tr>
<td>D</td>
<td>Multichain system, a detailed analysis</td>
<td>199</td>
</tr>
<tr>
<td>E</td>
<td>Cluster Distribution in SW fluids</td>
<td>203</td>
</tr>
<tr>
<td>E.0.0.1</td>
<td>Cluster Distribution</td>
<td>203</td>
</tr>
</tbody>
</table>

Bibliography | 207 |
List of Figures

2.1 Graphical illustration of physical aggregation mechanisms 15
4.1 Model's interaction potential 65
5.1 Energy histograms plot of a 32-mer chain 77
5.2 Specific heat capacity of a 32-mer chain for 2 independent runs 78
5.3 Internal energy of a 32-mer chain for 2 independent runs 79
5.4 Radius of gyration of a 32-mer chain 80
5.5 RMSD histogram plot of the 32-mer chain across the collapse transition 81
5.6 RMSD histogram plot of the 32-mer chain across the freezing transition 82
5.7 Average contact maps of a 32-mer chain across the collapse transition 83
5.8 Average contact maps of a 32-mer chain across the freezing transition 84
5.9 RMSD histogram of a 32-mer chain at the lowest temperature 85
5.10 RMSD histogram of contact maps a 32-mer chain at the lowest temperature 86
5.11 Average SASA of 32-mer at the lowest temperature 87
5.12 RMSC of a 32-mer chain at the lowest temperature 88
5.13 Average SASA of 32-mer at different temperatures 89
5.14 Energy histograms plot of 64-mer chain 95
5.15 Specific heat capacity plot of 64-mer chain for 2 independent runs 96
5.16 Internal energy of a 64-mer chain for 2 independent runs 97
5.17 Radius of gyration of a 64-mer chain 98
5.18 RMSD histograms plot of a 64-mer 98
5.19 Contact maps of a 64-mer across the collapse transition 99
5.20 Contact maps of a 64-mer across the second and third transitions 101
5.21 RMSD histograms plot of a 64-mer across the second-lowest temperature transition 101
5.22 Contact map of a 64-mer at the lowest temperature 102
5.23 RMSD histogram of a 64-mer at low temperature 103
5.24 Average SASA of the 64-mer residues at the lowest temperature 105
5.25 RMSC histogram of a 64-mer at the lowest temperature 106
5.26 RMSC histogram of a 64-mer across the intermediate-temperature transition 109
5.27 SASA of the low-temperature structure of the 32-mer and the 64-mer 111
5.28 Specific heat capacity of 64-mer, 128-mer, and 256-mer 115
5.29 Radius of gyration of 64-mer, 128-mer, and 256-mer 116
5.30 Specific heat capacity of 64-mer chain with different $\sigma/l$ 121
5.31 Diagram detailing the folding behavior of 64-mer chains with varying $\sigma/l$ 122
5.32 Specific heat capacity of 64-mer chains with $\sigma/l$ equal to 1.9 and 1.0 123
5.33 Radius of gyration of 64-mer chains with different $\sigma/l$ 124
5.34 $C_v, R_g, E$ of 64-mer chains with different $\sigma/l$ 125
5.35 $C_v, R_g, E^* \text{ of 64mer chains with different } \lambda$ 132
List of Tables

4.1 Symbolics, MJ potential, and type of 20 amino acids. 66
5.1 Simulation-temperature list of a 32-mer chain. 76
5.2 Simulation-temperature list of the 64-mer chain. 94
5.3 SASA analysis for the 64-mer at intermediate and transition temperatures. 107
5.4 Simulation-temperatures list of the 128-mer and 256-mer. 114
5.5 Summary of folding behavior of longer chains. 117
5.6 Number of transitions for chains with different ranges of interaction potential. 133
6.1 Temperature list used for the 5 protein sequences in isolation. 141
6.2 SASA analysis of the 5 isolated chains at $T^* = 0.033$. 141
6.3 Temperature of key transitions/properties of 5 optimized model-chains. 153
8.1 Properties of supercritical isotherms of an adhesive-sphere square-well fluid ($\lambda = 1.005$). 176
8.2 Critical temperature as a function of potential range. 178
C.1 Temperature list of the three 64-mer with $\lambda$ equal to 1.1 and $\sigma/l$ equal to 1.9, 1.6, and 1.3; and four 64-mer with $\lambda$ equal to 1.5 and $\sigma/l$ equal to 1.9, 1.6, 1.3, and 1.0. 198
List of Abbreviations

MJ  Miyazawa Jernigan
HP  Hydrophobic Polar
E.coli  Escherichia Coli
FAP  Familial Amyloid Polyneuropathy
E.coli  Escherichia Coli
rhGH  Recombinant Growth Hormone
CAB  Carbonic Anhydrase B
PFP  Protein Folding Problem
PSP  Protein Structure Prediction
NVT  canonical ensemble
NPT  isothermal-isobaric ensemble
NVE  microcanonical ensemble
MC  Monte Carlo
WHAM  Weighted Histogram Analysis
REM  Replica Exchange Method
MREM  Multiplexing variant of Replica Exchange Method
LES  Locally Enhanced Sampling
RE  Replica Exchange
REMUCA  Replica Exchange MULTicanonical Algorithm
MUCAREM  MULTicanonical Algorithm Replica Exchange Method
CEM  Constant Entropy Method
MREM  Multiplexing Variant Replica Exchange Method
MUCA  Multicanonical Algorithms
UNRES  UNited RESidues: a reduced model of polypeptide chains
BPTI  Bovin Pancreatic Trypsin Inhibitor
CI2  Chymotrypsin Inhibitor 2
DP  Discontinuous Potential
CP  Continuous Potential
SW  Square Well
PDB  Protein Data Bank
GBSA  Generalized Born Surface Area
DMD  Discontinuous Molecular Dynamics
RMSD  Root Mean Square Deviation
$R_g$  Radius Gyration
$C_v$  Specific heat capacity
SASA  Solvent Accessible Surface Area
high-T  high Temperature
low-T  Low Temperature
Physical Constants

Boltzmann’s constant \( k_B = 1.38064852 \times 10^{-23} \text{m}^2\text{kg}^{}\text{s}^{-2}\text{K}^{-1} \) (exact)
List of Symbols

\( E^* \) Dimensionless internal energy
\( RMSD \) Root mean square deviation
\( RMSD_{CM} \) Root mean square deviation in contacts
\( RMSC \) Root mean square deviation in contacts of the core
\( N \) number of beads (chain length)
\( X \) structural overlap parameter
\( k_i \) rate constant
\( c \) protein concentration
\( K \) equilibrium constant with \( K = k_i/k_{-i} \)
\( q \) order parameter measure of nativeness
\( E(q) \) thermal averaged effective-energy function
\( P(X_i) \) the probability of occurrence of the conformation described by the variables \( X_i \)
\( E(X_i) \) the energy of the conformation with probability \( P(X_i) \)
\( IR \) the ratio of strength of the side-chain’s hydrophobic interactions and backbone hydrogen bonding interactions
\( IR = \epsilon_{HP}/\epsilon_{HB} \)

\( k \) kelvin
\( mg \) milligram
\( ml \) milliliter
\( rpm \) rotation per minute
\( mM \) millimole
\( h \) hours
\( Kcal/mol \) kilocalorie per mole
\( \AA \) angstrom
\( ns \) nanosecond

\( \sigma \) hard sphere diameter
\( l \) bond length
\( \sigma/l \) stiffness, overlap parameter
\( \lambda \) range of interaction potential
\( \delta \) bond-length flexibility
Abstract

Doctor of Philosophy

Understanding biopharmaceutical aggregation using minimalist models based on square-well potential

by Hamza J. MAGNIER

November 4, 2016
Protein misfolding and aggregation are the cause of many problems within the biopharmaceutical industry and medical fields. Although many experimental studies have been implemented in vivo in order to understand this process, the mechanism occurs in time and length scales inaccessible to conventional experiments. On the other hand, computational studies have shown significant improvement in elucidating key aspects of the aggregation pathways and gain insights to the folding behavior of the proteins. Consequently, this makes computational modeling an ideal complement to experiment in understanding the generic behavior and mechanisms of aggregation. This study is concerned with DynamO, a coarse-grained, off-lattice, general event-driven discontinuous molecular-dynamics simulation package. This simulator offers a unique opportunity to gain insight into the process of protein aggregation by displaying the optimal $O(N)$ asymptotic scaling of the computational cost with the number of particles $N$, rather than $O(N \log N)$ scaling found in most standard algorithms.

The study was split into two loosely related projects: in the first project, a computer model was developed in which the effect of model parameters on folding behavior and characteristics of isolated peptides is investigated. The model parameters include chain stiffness (an overlap parameter defined as the ratio of the hard-core diameter to bond length $\sigma/l$), range of interaction potential $\lambda$, sequence, and chains length $N$. Based on the model chosen from systems of isolated chains, aggregation in multichain systems is studied. In another project, we simulate various square-well fluid systems with different ranges of interaction potential in order to understand the phase behavior of proteins due to its relevance to aggregation and many bioprocessing events.

Changing the model parameters shows different folding behaviors. The model-chains with 64 residues, $\lambda$ equal to 1.1 and $\sigma/l$ equal to 1.9 is the least computationally expensive model displaying all the characteristics found in real proteins. We introduce a new order parameter which divides the conformational space into folded and unfolded ensemble-structures, this order parameter corresponds to a transition in the folding behavior of the chains. We define a native state ensemble as an ensemble of structures with small deviation in contact maps for spheres inaccessible to the solvent defined as the core of the chain. This native ensemble corresponds to the structures exhibiting low-temperature fluctuations simulating the ‘breathing motions’ of real proteins which is considered responsible for their catalytic activities. On the other hand, the non-native ensemble unfolds at higher temperatures, which increases the propensity for aggregation by forming intermolecular contacts, and therefore reproduce the behavior of proteins under severe solution conditions which occurs in bioprocessing (this includes high concentration, temperature, pressure, pH ...). The behavior of multichain systems shows that it is possible to correlate the aggregation propensity of chains at room temperature from the behavior of chains in isolated system at the collapse temperature, which in turn correlate with the stability of the low-T ensemble.

In the second project, we developed a more efficient way of calculating the critical temperature in SW fluids even for strongly short-ranged systems which are especially difficult to simulate. In the supercritical region, every isotherm obeys the linear equation for the pressure with a high precision within the bounds of uncertainty. The linear equation $p_m = p_0 + R_m \rho$ with $R_m$ being the constant isothermal rigidity $(dp/d\rho)_T$. The constant rigidity can be used to estimate directly a critical temperature ($T_c$) and critical pressure ($p_c$), respectively, and also to obtain the pressures and densities of the percolation loci based on an empirical quadratic nature of change in pressure with densities outside the percolation loci. Identifying the critical temperature and how it depends on the pair potential is very important in formulations with a growing need to predict when the solution will go opalescent.
Declaration of Authorship

I, Hamza J. MAGNIER, declare that this thesis titled, “Understanding biopharmaceutical aggregation using minimalist models based on square-well potential” and the work presented in it are my own.

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright Statement

• The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

• Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

• The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

• Further information on the conditions under which disclosure, publication and commercialization of this thesis, the Copyright and any Intellectual Property University IP Policy (see http://documents.manchester.ac.uk/display.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University’s policy on Presentation of Theses.
For the owner of Era and time, same as his fathers... For The Leader of the resplendent and Master of the guardians.

“Mountains may move from position, but you should not move from yours. Grit your teeth. Lend to God your head. Plant your feet firmly on the ground. Set your eyes on the remotest foe and close your eyes. And be sure that success is only from God.”

Imam Ali Ibn Abi Talib

“Our Big Bang to a well-ordered universe, Entropy - which is rather unknown to man - is nothing but the ultimate proof ...”
Acknowledgements

First and far-most, I would like to thank my supervisor Dr. Robin Curtis for the support, the help and the guidance provided throughout my research time. Dr. Robin was a supportive and insightful person who did not hesitate to lend me a hand in time of need.

Special thanks goes to Professor Leslie Woodcock for giving me a new perspective and a deeper appreciation of the wonders of science.

Many thanks goes to Dr. Leo Lue and Dr. Marcus Bannerman for all the technical help which was invaluable to completing my studies.

Finally, I would like to thank my family and friends for all the love provided in the time of needs. Special thanks to my mother, whom without her I would not have the strive to fulfill all ambitions.
To my mother .......
Chapter 1

Introduction

Protein aggregation involves the intermolecular association of partially-unfolded or misfolded proteins into large agglomerates (Das et al., 2014). This is the most common and troubling manifestations in both of biotechnological, medical fields, and food industries (Koo, Lansbury, and Kelly, 1999). The interest in understanding and predicting aggregation has been increasing significantly in time. For the past two decades the number of candidate-biopharmaceuticals has grown rapidly, fueled by technological advancements, and high-throughout methodologies. Despite the effort, rapid commercialization of protein drugs has not been realized, due to various instabilities during the process, aggregation being undoubtedly the most disturbing. Aggregation occurs at almost every stages of the production process, including fermentation, refolding formulation, purification and storage (Wang, 1999; Wang, 2005). There have been rapid advances in the upstream processing of protein therapeutics, which has shifted the bottleneck to downstream purification and formulation. This begs the need for exploring means to develop improved predictive ability that could then be used for a better understanding of how protein-protein interactions depend on formulation conditions (pH, ionic strength, buffer type, presence of excipients) and how these impacts upon the initial steps in protein self-association and aggregation. Aggregation occurs at a large range of length and time scales, which makes part of the process inaccessible to experimental study and observations. This limitation makes computational modeling an ideal complement to experiments in simulating the generic features of the aggregation process, especially for examining the initial steps in aggregation.

Studies on aggregation branch into two major categories: aggregation by natively unfolded
proteins and aggregation from the native state. This study is mainly concerned with aggre-
gation from the native state, while most simulation studies have focused on aggregation of
natively unfolded proteins.

Aggregation by natively unstructured proteins, commonly known as amyloid aggregation
or fibrillar aggregation, is directly implicated in the pathology of more than twenty neu-
rodegenerative diseases including Alzheimer’s, Parkinson’s, Huntington’s and prion dis-
eases (Koo, Lansbury, and Kelly, 1999; Pande et al., 2001; Hall and Waggoner, 2006). In each
of these disorders, known as amyloidosis, a specific protein slowly accumulates in fibril-
lar tangles or plaques, destroying the architecture and function of the surrounding tissues
which leads to degenerative and ultimately fatal consequences. This type of aggregation has
been extensively studied both in vivo and in vitro. Many groups studied amyloid aggregation
from initially unfolded states using various models. This includes three-dimensional lattice
models with Monte Carlo simulations, Go-models, HP models, and many others (Hall et al.,
2006) with seminal studies on amyloid aggregation done by Hall et al. (Hall and Waggoner,
2006; Nguyen and Hall, 2006) using intermediate coarse-grained models.

The second major category of aggregation is that occurring from the native state. Aggrega-
tion from the native state is more relevant for bioprocessing and formulation applications. It
occurs throughout various unit operations when protein solutions are under many forms of
stress. Moreover, finding liquid formulations with shelf lives of up to two years is increas-
ingly difficult for some of the newer therapeutics, which have been engineered for activity,
but their formulations are often viscous, can phase separate, and have a high propensity for
irreversible aggregation. A good example is the refolding from inclusion bodies during pro-
cessing (Agrawal et al., 2011). In inclusion bodies, proteins are aggregated in native states.
Obtaining natively folded proteins requires first solubilizing the proteins under chemically
denaturing conditions and then refolding the protein by removing the denaturant. There-
fore extensive studies are concerned with the competition between aggregation and folding
in the native state (Bratko and Blanch, 2001; Blanch et al., 2002; Bratko and Blanch, 2003;
Bratko et al., 2006; Cellmer et al., 2005c; Bratko et al., 2007; Cellmer et al., 2005b; Cellmer
et al., 2005a).

Another example of problematic biopharmaceutical aggregation is the association of Insulin
during formulation to form oligomers via the native state (Agrawal et al., 2011) and aggregation during storage and shipping of protein drugs (Chi et al., 2003). Recent work correlated the aggregation propensity of chains as a function of concentration and temperature (Woodard, Dunatunga, and Shakhnovich, 2016). In another study, aggregation-prone regions were identified in monoclonal antibodies from their native states (Chennamsetty et al., 2009). Although this area of aggregation is slightly studied in literature, the pioneer work of Bratko and Cellmer group on this topic is used as a benchmark in this study. After characterizing the behavior of chains in isolation, the competition between folding and aggregation was studied using multichain simulations (Bratko et al., 2007; Cellmer et al., 2005c). The model used was a three-dimensional on-lattice model to examine the refolding problem. A multichain system consisting of 64 monomer chains was simulated to study the role of interprotein and intraprotein interaction on the aggregation propensity (Bratko et al., 2007; Cellmer et al., 2005c). Another study used an off-lattice model of three 46 monomer chains to investigate the effect of folding enhancers, or chaperons, on the refolding yield. It was suggested that the chaperon prevented partially folded species from forming aggregates (Bratko and Blanch, 2001). The model used was originally developed by Honeycutt and Thirumalai (Honeycutt and Thirumalai, 1992), and combined with a new interaction-energy scale as an extension to the Miyazawa-Jernigan (MJ) potential developed by Leonhard et al. (Leonhard, Prausnitz, and Radke, 2003a). Cellmer and coworkers used Langevin simulations to study the competition between folding and aggregation using minimalist off-lattice model protein using a hydrophobic-polar (HP) potential (Cellmer et al., 2005b). The results suggest that multiple mechanisms for aggregation occur, and that the association process is much faster than folding. Moreover, a three-dimensional on-lattice model was used to study the competition between intramolecular and intermolecular interactions starting with initially unfolded states leading to folded and misfolded chains (Bratko and Blanch, 2001).

As a first step towards understanding protein aggregation, a native state - in which the protein is biologically active - needs to be defined. Two definitions exist in literature, the first one is defining the native state as a singular conformation at the global minimum of the complex free-energy landscape (Kaya and Chan, 2003; Bratko and Blanch, 2001), while the second view suggest that the protein is biologically active in a range of conformationally
fluctuating structures. The native structures are rather a Boltzmann ensemble of conforma-
tions at low temperature (Tang and Dill, 1998).

Defining the native state as a stagnant structure was used extensively in investigating the
kinetics and thermodynamics of protein folding. Many extended this definition to study
aggregation (Bratko et al., 2007). This definition is rather appealing, specially in simplified
models like on-lattice HP models which allows determining the lowest energy state by exact
enumeration of fully compact conformations (Sali, Shakhnovich, and Karplus, 1994).

While this definition shows invaluable benefits, it does not take into account conformational
fluctuations of the chain at low temperatures. Real proteins at room temperature exhibit a
‘breathing-motion’ while maintaining their biological activity. These conformational fluctu-
ations are sometimes essential for the catalytic function of globular proteins and enzymes
function, induced-fit mechanisms of ligand-binding (Koshland, 1958; Koshland, 1959), and
for allosteric regulations (Tang and Dill, 1998). X-ray diffraction experiments on ribonu-
clease A detected protein motion (structure fluctuations) at different temperatures between 98k
and 320K (Tilton, Dewan, and Petsko, 1992). This structure fluctuation is in agreement with
the breathing-motion of proteins. If proteins needs ‘flexibility’ for their function then the
native structures are the ensemble of all fluctuating conformations that are sampled in this
motion. Accordingly, defining the native state as a single conformation is rather simplistic
while defining native conformations as an ensemble of fluctuating structures appears to be
a more realistic description of real proteins.

Considering the native states as an ensemble of fluctuating conformations means that any
structure should be able to take exclusive part of one of the two ensemble: either the folded
ensemble (with native conformational fluctuations) or the unfolded ensemble. This begs the
need to differentiate these ensembles based on a given property. Some attempts were made
to find this differentiating characteristic. One of these attempts is the pioneering work of
Tang and Dill (Tang and Dill, 1998) on computational modeling of protein behavior. Tang
and Dill refer to a conformational temperature below which large conformational fluctu-
ations are frozen-out, and define the native ensemble as the structures below this confor-
national temperature. Studies on myoglobin (Keller and Debrunner, 1980), ribonuclease
A (Tilton, Dewan, and Petsko, 1992), cytochrome c complex (Noeck et al., 1991), crambin
Chapter 1. Introduction

(Usha and Wittbort, 1989), superoxide dismutase (Andreani et al., 1995; Melchionna, Falconi, and Desideri, 1998) and bacteriorhodopsin (Ferrand et al., 1993) have defined a temperature around 200K below which proteins exhibit small motions, and above which the motion is significantly larger. All these studies suggests that certain motions are frozen-out at low temperatures, but the exact nature of these motions is still out for interpretation. In a different approach, Cellmer et al. (Cellmer et al., 2005b) calculated a chain’s free energy for several values of the structural overlap function at the folding transition temperature. The structural overlap function ($X$) is a measure of a snapshot’s similarity to the lowest energy structure. The minimum at $X$ equal to 0.55 corresponds to compact non-native states. The local maximum at $X = 0.35$ is the thermodynamic barrier to folding. Thus, individual chains that have overcome this barrier, i.e. those with $X$ values less than 0.35, are in the native state. In another study, they refer to a local energy maximum when calculating the free energy as a function of potential energy. Accordingly, any chain that overcomes this energy barrier is a folded chain (Cellmer et al., 2005c).

We direct our attention towards the behavior of the core of protein chains (group of beads inaccessible by the solvent molecules, buried inside the chain) driven by the general consensus that it plays a major role in proteins folding, unfolding, and aggregation (Dobson, Sali, and Karplus, 1998; Dinner et al., 2000).

Looking first at aggregation mechanisms, the old view suggests that aggregation is caused by association of totally unfolded chains (Dill et al., 1995; Stiger and Dill, 1993). Studies that followed have shown that aggregation is rather driven by association of partially unfolded or folded intermediates (Nagpal et al., 2015; Fink, 1998; Fields et al., 1992). Bratko et al. suggest that the degree of interprotein association strongly correlates with the exposure of the hydrophobic core as proteins unfold (Bratko et al., 2007; Leonhard, Prausnitz, and Radke, 2003b; Broglia et al., 1998). When two attractive sites of the exposed core form a contact, irreversible aggregation occurs. Moreover, the core contributes largely to the folding and unfolding phenomena. Folding of the chain is driven by the collapse of the hydrophobic core which sacrifices conformational entropy for an increased potential energy (Cellmer et al., 2005b). This also presents a solution to the known ‘Levinthal’s paradox’ (Miller et al., 1992) which addresses the quick folding of proteins in a single structure out of the infinite amount of available conformations.
Moreover, if the core is responsible for both small conformational fluctuations simulating the protein breathing motions, and large conformational fluctuations inducing unfolding and aggregation, there is a strong correlation between aggregation propensity and the folding temperature implying the behavior of the protein at high temperatures correlates with the properties at room temperature. The main point is therefore to determine whether or not the conformational fluctuations of the native state ensemble correlate with the free energy difference between the native ensemble and the unfolded state. This is based on the experimental observation that aggregation, which is believed to occur through non-native states, correlates with the melting temperature of the protein, which in turn reflects the free energy difference between the folded and unfolded ensembles. Another possibility could be that aggregation occurs through the native ensemble.

In summary, at low temperatures chains exhibit small conformational fluctuations or ‘breathing motions’. At high temperatures, unfolded chains collapse into compact structures due to the formation of a hydrophobic core. Moreover, at slightly high temperature or a change in solution conditions, many chains partially unfold and aggregate by exposure of their attractive cores. There is a strong experimental correlation between aggregation propensity and the folding temperature implying that the aggregation occurs through the unfolded state, which is not expected to be significantly populated at room temperature. In this report, we aim at answering the following questions: what is the nature of conformational fluctuations at low temperature? How do we differentiate between partial unfolding, which promotes aggregation, and conformational fluctuations, which are important for protein activity? Understanding this behavior requires mapping the reduced temperatures in modeling experiments to physical temperature in the real world. What is the role of the core in folding behavior and aggregation propensity of protein chains? Can we use the properties of the core to define a native-ensemble of states, and hence divide all structures in the folding pathway between folded and unfolded allowing us to calculate energy difference between states and determining whether or not aggregation occurs though unfolded or folded states? Defining a native state ensemble will also allow determining whether or not aggregation occurs though unfolded or folded states.

Here we introduce a new off-lattice coarse-grained model to study native-state aggregation. The model corresponds to a chain of spheres interacting through square-well potentials.
The simulator *DynamO* (Bannerman, Sargant, and Lue, 2011), which employs the discontinuous molecular dynamics algorithm, is used for the simulation of chains with up to 256 monomers and a multichain system of more than 8 polymers.

A brief outline of the report is as follows. Chapter 2 outlines a background study concerning simulations of protein folding and aggregation pathways. This includes a review of molecular simulation algorithms and potential models used for studying protein aggregation. Special attention is given to the work most relevant to this project, mainly discontinuous molecular dynamics and the development of the quasi-chemical Miyazawa-Jernigan approximation of the residue potential energies. Chapter 3 articulates the hypothesis of the project and the specific objectives of the study. Chapter 4 introduces the model-chain used in the simulation, along with the methodology implemented. It also describes the set of various analysis tools and their implementation. Chapters 5, 6, 7 and 8 cover the main results and discussions of this project. Chapter 5 includes simulation of isolated chains and the effect of model parameter. We first looked at the change in folding behavior of isolated chains by varying the length of the chain (equivalent to number of beads $N$), the ratio of the hard sphere diameter to bond length termed the overlap parameter ($\sigma/l$), and the range of interaction-potentials ($\lambda$). This chapter allowed us to identify models that exhibit the generic behavior of the folding of ‘real’ proteins. In chapter 6 we present simulation results for 5 isolated model-chains with the set parameter chosen previously but with varying sequences. These will be used later-on as a reference in multichain systems. We also introduce a new definition of the native ensemble based on the order parameter defined as the fluctuations in the contact maps formed by spheres contained within the protein core. Comparing the data of the chains gives insight on the correlation between properties at low temperature and behavior at the folding temperature. This definition allows labeling each structure in the simulation as folded or unfolded. In chapter 7, we present results for 5 multichain systems (one for each chain simulated in isolation) up to 8 model-chains per simulation box. We examine the possibility of estimating the aggregation propensity of protein chains from their behavior in isolation at room temperature. In chapter 8, we introduce results of square-well fluid simulations and a new method for calculating critical parameters based on empirical observations. The new phenomenological description of the phase behavior draw an analogy between protein solution and systems of bulk fluid. Much of the work in the area of
proteins crystallization focused on the connection between protein-protein interactions and the phase diagram by revisiting critical opalescence in the light of the new interpretation of the phase diagram. Chapter 9 synthesizes the information from the results chapter. Finally, Chapter 10 provides suggestions for future work, and outlines the next objectives building on this work.
Chapter 2

Background

2.1 Protein folding Problem

Proteins, formed by a specific sequence of amino acids, are involved in virtually every biological process. Their function ranges from catalysis of chemical reactions to conservation of chemical potential across the cell membrane of living organisms (Dobson, Sali, and Karplus, 1998). These linear chains of 20 different naturally occurring amino-acid blocks are synthesized on ribosomes in a specific order from information encoded within the cellular DNA depending on their ultimate function. Following their biosynthesis, most proteins fold into tightly compact unique three-dimensional structures. This specific structure is responsible of their activity and is commonly referred to as the native state (Dobson, 2004). Due to the substantially large number of possible conformations for a single chain with the average size of those in our bodies, the folding of these chains into unique structures is recognized as a perplexing question. So how do these proteins find their way into this distinct three-dimensional conformation?

2.1.1 Protein folding: a thermodynamic problem

The protein folding problem lies at the heart of molecular and cellular biology. Experimental and computational studies contributed a lot in understanding the process through which an extended polymer always folds into a native three-dimensional structural shape (Colombo and Micheletti, 2006). As the number of possible conformations for a polypeptide chain is very large, a systematic search for the native structure would require an almost infinite...
length of time. Natural proteins exhibit characteristics and properties which are not typical of random polymers. Rather, it is a very special group of molecules with the ability to fold to a unique structure in very short time and hence generate enormous selectivity and diversity in their functions.

The publication of Anfinsen’s results in 1973 (Anfinsen, 1973) marked the conceptualization of the protein folding problem. His work on ribonuclease revolutionized the theory of ‘The thermodynamic hypothesis’, which states that the native folded state corresponds to the global minimum free energy of the system at the physiological conditions. Anfinsen showed that the refolding of ribonuclease after denaturation by reductive cleavage of its total four disulfide bonds will always find the exact 1 of the 105 possible pairings of eight sulfhydryl groups to form these disulfide linkages. This elegant process of natural selection is claimed to be driven only by free energy of conformation that is gained in going to the stable unique native structure. Anfinsen therefore categorizes the folding process as a thermodynamic problem, with slight indication to the kinetic discrepancy in time to fold a chain of 124 amino-acid residues between \textit{in vitro} (few hours), and \textit{in vivo} (2 minutes). Anfinsen explained this inconsistency by the existence of biological catalytic enzymes which drives the process rapidly in the endoplasmic reticulum of the cells. The publication of these results sets the scene for decades of research dedicated to determining how a sequence comprising just 20 different amino acid could dictate millions of complex three-dimensional shapes (Anfinsen, 1973).

2.1.2 Protein folding: a kinetic problem

In 1968 Levinthal explained the impossibility of observing high folding rates if the proteins were to randomly sample all available conformations of the polypeptide chains (Zwanzig, Szabo, and Bagchi, 1992). Levinthal argued that the fast kinetics can be explained by the existence of energetic and entropic pathways driving the folding process. This argument stemmed from the idea that in the absence of such forces, the native state is reached only by random search of the available conformational space. The time required for such search is incompatible with the actual time of folding in real proteins (of the order of milliseconds) even for very small peptides. This was later-on known as the ‘Levinthal’s paradox’ (Zwanzig, Szabo, and Bagchi, 1992; Dill and Chan, 1997).
As an answer to this paradox, Levinthal explained that the protein not only achieves the global energy minima, but does so quickly. These two mutually exclusive criteria were known to be the thermodynamic and the kinetic control: thermodynamic control refers to the fact that the protein’s native state is at the global free energy minima and is reached by the unfolded protein independently of the folding path (i.e the initial denaturing condition). This requires long extensive exploration of conformational space and costs a lot of time. The kinetic control means that proteins fold in biological timescale (i.e very quickly). This depends on the folding pathway, the initial conditions of denaturation, and the existence of local energy minima that might prevent the protein from reaching the ground state. These observations led to the conclusion that proteins fold by specific ‘folding pathways’ (Dill and Chan, 1997).

2.1.3 Protein folding: a hydrophobic core

In 1992, Zwanzig and coworkers (Zwanzig, Szabo, and Bagchi, 1992) showed that a small and reasonable energy bias introduced against locally unfavorable configurations can reduce Levinthal’s time to a biologically significant size in a simple model. This energy bias was later-on correlated with a group of very strong contacts between a small number of amino acids, collectively known as the ‘core’, the ‘folding core’, or the ‘nucleus’ (Dinner et al., 2000). The formation of such contacts provides a solution to Levinthal’s paradox by increasing the reaction rate and simplifying the search process. Chymotrypsin inhibitor 2 (CI2) is a well documented protein which displays the formation of a core and has been extensively studied both experimentally and using computer simulations. A disturbance in the core is seen to cause a change in the folding pathway. The folding pattern changes from three-state to two-state folding due to a single mutation in the buried residues (Khorasanizadeh, Peters, and Roder, 1996). Similarly, Jackson (Jackson, 1998) argue that a single substitution in core residues drive a significant difference in the rate of folding according to protein engineering studies. The case of the four-helical bundle protein ‘ROP’ reports a three times increase in magnitude of the folding rates by a change in core residues without affecting the native structure. Computer simulation also reports the significant importance of the core for 125-mer model chains (Dobson, Sali, and Karplus, 1998). The nucleus, consisting of 30 residues or so, is formed early during the folding process, and the remainder of
the polypeptide chain condenses rapidly around it. Accordingly, the core of proteins plays an important role in the folding behavior of chains.

### 2.2 Protein misfolding and aggregation

#### 2.2.1 Aggregation in bioprocessing

Aggregation in bioprocessing is the most relevant subject to this work, as most of aggregation phenomena occur from the native state. This type of aggregation is much less studied compared to amyloid aggregation or aggregation from non-native states, which is extremely important in medicine. Protein aggregation is induced at almost every stage of biopharmaceutical processes. This includes protein refolding, fermentation, shearing/shaking, freeze-thawing, drying, reconstitution, and storage (Wang, 1999). It is one of the largest contributors to reduced yields in bioprocessing, which hinders rapid commercialization of protein drug candidates despite the explosive growth in biopharmaceutics in the past couple of decades (LeTreut, Képès, and Orland, 2016; Ratanji et al., 2014; Wang, 1999). Aggregates usually trap the protein in an inactive and unrecoverable state.

During formulation, many factors contribute to protein aggregation. Increasing protein concentration increases the rate of aggregation (Wang, 2005). After cell culture, the protein solution is purified to remove viruses, impurities, even host cell. The solution is subjected to a wide range of ionic strength, pH, and fluctuating protein concentrations which leads to an increase in aggregation (Andya, Hsu, and Shire, 2003). Moreover, the salt type used in controlling the ionic strength plays a major role in the aggregation propensity of the solution. Temperature changes can be a cause of protein aggregation. One of the methods of protein purification is heat treatment at 60 °C. Many proteins aggregate during this process, unless the solution is well protected. Spray drying a solution of 10 mg/ml rhMAb is reported to aggregate by 5.6% (Andya et al., 1999). Another viral inactivation method is the freezing of protein solutions, in which case aggregation occurs due to formation of ice-water interface, phase separation, and potential pH change. Removal of parts of the hydration layers during the drying process (the water content of a dried protein product is usually less than 10%) may alter the native state of a protein and promote aggregation. Freeze thawing a 50 mg/ml
solution of hemoglobin in PBS for 5 times causes the formation of $1.4 \times 10^4 \text{ ml}^{-1}$ particles. In this case the solution was held at a temperature of -20 °C and a pH of 7.4 (Kerwin et al., 1998).

Other miscellaneous processes contribute to solution aggregation. During protein shipping and storage (short- and long- term storage, 18-month shelf lives, etc.) the solution may suffer temperature changes, and induced thermal aggregation may occur. Although low-temperature storage is favorable, this may also cause aggregation (similar to aggregation during the freezing process). Maintaining ambient-temperature conditions requires extensive optimization of the protein immediate environment. Shaking and shearing during separation or storage processes is reported to induce aggregation. The alignment of proteins at the air/water interface maximizes the exposure of hydrophobic residues to air and promotes aggregation. Agitation at 100 rpm of a 0.6 mg/ml insulin solution at pH 7.4 yields aggregates in 2 days (Sluzky, Klibanov, and Langer, 1992). Similarly, shaking a solution of 1 mg/ml FXIII protein at 400 rpm and 23°C for 24 h in 0.1 mM EDTA and 10 mM Tris at pH 8 causes the solution to fully aggregate (Kreilgaard et al., 1998a; Kreilgaard et al., 1998b).

Some cases of aggregation from the non-native state also occur during cell growth. For example, aggregation may occur during cell culture due to the accumulation of high amount of proteins (Cromwell, Hilario, and Jacobson, 2006). A well-known example is the formation of amorphous aggregates, which are referred to as inclusion bodies, from recombinant proteins over-expressed in E.coli (Cellmer et al., 2005b). These aggregates need to be refolded at the end of the process to recover their biological activity. However, refolding of multiple proteins at moderate to high concentrations is prone to off-pathways aggregation, which decreases the yield of properly folded proteins. This is economically challenging as it leads to high production costs associated with large buffer volumes and increased tank sizes. Due to the lack of understanding of the aggregation process and mechanisms, the end results of the bio-industry formulations are still unsatisfactory, despite all methods used to prevent, or at least inhibit, aggregation (Cellmer et al., 2005b).

In short, aggregation in drug formulation does not only cause a decrease in the yield of production and hence an increase in the production cost, but the presence of aggregates in the final protein formulation is unacceptable as the folding of aggregates often elicit an immune response in the patient, which is prevents product release (Wang, 2005).
2.2.2 Aggregation in human etiology

More than twenty human diseases are associated with protein aggregation (Wang, 2005). Specifically, abnormal protein aggregation is thought to correlate with most, if not all neurodegenerative diseases. These include Alzheimer’s (AD), Parkinson’s, Creutzfeldt-Jakob, motor neuron, the large group of polyglutamine disorders, including Huntington’s disease, as well as diseases of peripheral tissue like familial amyloid polyneuropathy (FAP) (Koo, Lansbury, and Kelly, 1999), cataracts caused by protein crystallization (Pande et al., 2001), Down’s syndrome, prion diseases, sickle-cell disease (Curtis and Lue, 2006) and Bovine Spongiform Encephalopathy known as the mad cow disease.

2.2.3 Aggregation types

There are two general categories of protein aggregation: chemical aggregation and physical aggregation. This project is concerned only with the latter one. A brief definition is introduced for each of the two types without being exhaustive. Proteins may physically associate with each other without any changes in their primary structure, which often leads to the formation of soluble oligomers. On the other hand, in chemical aggregation, partial protein unfolding often leads to the formation of insoluble protein aggregates, which are irreversibly formed and difficult to recover. The latter process can also be brought about by the formation of interprotein covalent bonds, which cross-link proteins together. Chemically-induced aggregation includes disulfide bond formation/exchange, non-disulfide cross-linking pathways, oxidation, Maillard reaction, and others (Wang, 2005).

Both mechanisms are not exclusive; i.e. they can either occur individually or simultaneously, with each type leading to a different end result. One example is the chemical aggregation of insulin leading to the formation of either soluble dimmers via cyclic anhydride intermediate or insoluble disulfide-bonded aggregates. Physical aggregation processes lead to the formation of either insoluble fibrils or soluble hexamers (Sluzky, Klibanov, and Langer, 1992; Costantino, Langer, and Klibanov, 1994; Darrington and Anderson, 1995).
2.2.4 Physical-aggregation mechanisms

Aggregation pathways can be separated into five general categories: (1) aggregation by unfolding/folding intermediates and states, (2) aggregation by native self-association, (3) aggregation from the unfolded states, (4) aggregation through direct chemical linkages, and (5) aggregation through chemical degradation. The first three types of aggregation are physically-induced. Although this report focuses on physical aggregation, we will briefly define all types of aggregation with special focus on the relevant ones.

Figure 2.1: A simplified illustration of aggregation pathways: N = native structures, I = intermediates structures, U = unfolded structures, A = aggregates and P = precipitates. We label the major aggregation pathways as (1) aggregation through unfolding intermediates and unfolded states (pathway 1), aggregation through native protein self-association (pathway 2) or unfolded protein self association (pathway 3), and aggregation through native and unfolded chemical linkages (pathway 4).

Figure 2.1 shown the three major aggregation pathways: (1) aggregation through unfolded and folded intermediates states (pathway 1), aggregation through native protein self-association (pathway 2) or unfolded protein self association (pathway 3), and aggregation through native and unfolded chemical linkages (pathway 4). Aggregation through chemical degradation is not shown here.
Intermediate states, which are partially unfolded/folded states, lead to the formation of aggregates (I to A). The definition of these newly formed aggregates are quite strenuous. This definition needs to differentiate between the aggregated structures, intermediate structures, and precipitates (P). Wang and coworkers (Wang, Nema, and Teagarden, 2010) defined the newly aggregates as structures that do not have a native conformation, and are at least twice as big as the completely folded states (N). If one of the two conditions are not satisfied, then such structures cannot be considered as aggregated. For example, if dimers or trimers exhibit native conformation, they cannot be considered as aggregates despite having a size equal or greater than twice that of the native protein. Nonetheless, newly-formed aggregates are soluble within a certain range in size and solvents solubility limits. Once these limits are exceeded, they precipitate into insoluble aggregates (A to P). The insoluble aggregates follow either an ordered pathway (fibril formation) or a disordered pathway (amorphous precipitation). Pathway (2) shows aggregation occurring through native-state association, which is not very common. Indeed, it is not clear whether or not this pathways actually occurs. This depends on the definition of the native state and native ensemble investigated in details in chapter 6.

This definition of native state influences the definition of intermediate state (I). Usually, native protein chains in a solution are in equilibrium with small amounts of unfolded intermediates under normal conditions. These intermediates can be either native-like intermediates (NI), also referred to as molten globules, or non-native like intermediates (NNI). NNI occur if the degree of unfolding is large (Wang, Nema, and Teagarden, 2010). The molten-globule state is of special interest in this work, due to the role of the hydrophobic core in the folding process. Although the old view suggests that aggregation is induced by unfolded protein structures (Dill et al., 1995; Stiger and Dill, 1993), recent studies suggest that folded and unfolded intermediates both play a major role in driving the aggregation process regardless of their instability and low concentrations (Nagpal et al., 2015; Das et al., 2014; Fink, 1998; Fields et al., 1992). Studies on scrapie amyloid protein, or prion protein (Safar et al., 1994), recombinant human growth hormone (rhGH) (Bam, Cleland, and Randolph, 1996), insulin (Brange et al., 1997), carbonic anhydrase B (CAB) (Cleland et al., 1992; Hammarstrom et al., 1999), and phage P22 wild-type coat protein and its mutants (A108V, G232D, F353L) (Teschke, 1999) present an overwhelming evidence to that matter. The role of intermediates
in initiating the aggregation process is also supported by computer experiments (Nagpal et al., 2015; Gupta, Hall, and Voegler, 1998; Istrail, Schwartz, and King, 1999). On the other hand, totally folded and unfolded chains are actually less susceptible to aggregate because the hydrophobic core is either totally buried inside the chain (i.e., not in contact with solvent molecules), or completely scattered (Uversky et al., 1999; Zhang, Lu, and Liu, 2008). What initiates aggregation are rather patches of continuous hydrophobic groups.

Intermediates might form from conformational fluctuations of the native states. Thus the question that arises is when would this modified conformation fall within the native-like ensemble? Also, at which point are certain structural deviation from the native conformation considered as non-native-like structure? These questions and more will be revisited in the following chapters as we differentiate native and non-native structures.

The second type of physical aggregation is aggregation through self association and follow paths denoted as pathways 2 in figure 2.1. In this mechanism native proteins form aggregates directly without going through intermediates (N to A). It is not clear whether this pathway actually occurs, but Wang et al. propose that such transitions are possibly induced by electrostatic interaction, hydrophobic interactions and other weak forces such as Van der Waals forces. The reversible aggregates A are the precursors of the non-reversible precipitate (P). Another possibility is that self association of native states leads to reversible precipitation (P) or crystallization where the protein is still in the native state. This pathway does not appear in figure 2.1. The transitions between N and A are often accompanied with conformational changes. The tendency of protein’s self association is indicated usually by the osmotic second virial coefficient which reflects the colloidal stability and instability of the chains: positive $B_{22}$ indicates protein-protein repulsion and favors protein-solvent interaction, while a negative $B_{22}$ show that protein-protein attraction are favored over protein-solvent interaction, and proteins exhibit a colloidal instability (Lford et al., 2008).

The third aggregation pathway is aggregation from the denatured states as shown in figure 2.1 (3). Proteins that are naturally unfolded or exhibit one step folding follow this mechanism (Wang, Nema, and Teagarden, 2010). This type of physical aggregation (U to A to P) is not of interest in this report as we are looking at aggregation from the native state.

Protein chains can be directly cross-linked by disulfide bond formation and/or exchange.
Since this type of chemical degradation is promoted by sulfide linkages, the presence of cysteines amino acids outside the buried core, i.e on the surface of the chain, will increase aggregation formation in chains. Other non-disulfide cross linkages has also been reported but not as frequently as the first type. These include Maillard type reactions, di-tyrosine formation, and oxidation.

The last type of chemical aggregation is the chemical degradation-induced protein aggregation. In this pathway, the properties of the chains change. This includes changes in secondary and tertiary structures, hydrophobicity, thermodynamic and kinetic energy barriers to folding and unfolding, and many more.

In this report, we are studying which pathways and intermediate is being followed in the aggregation process, whether the intermediate belongs to the native-state, unfolded-state, or a partially folded state. Thus the key point is to define the native folded ensemble and unfolded ensemble as the partially folded state could be considered as part of either ensembles.

### 2.2.5 Kinetics of aggregation

The kinetics of aggregation depend on the slowest (rate-limiting) step in the pathways. In many instances, the rate limiting step is the partial unfolding of the protein. In this case, the aggregation kinetics would be first order in the concentration of the native protein, that is \( k_i c \) where \( c \) is protein concentration. Most methods for preventing protein aggregation target this step of the aggregation process. Additives such as sugars, salts or polymers are included in the solution to either decrease the value of \( k_i \) or to stabilize the folded protein by decreasing the equilibrium constant \( K_i = k_i/k_{-i} \). Alternatively, the rate limiting step can be the formation of a critically sized nucleus as in protein crystallization. In this case, an activation barrier to aggregation exists leading to an exponential dependence on protein concentration and a lag time preceding the formation of aggregates. This type of aggregation is often observed when the precipitated aggregates form long ordered structures termed fibrils (Weiss, Young, and Roberts, 2009). Alternatively, when amorphous aggregates are formed, the rate determining step is set by the partial unfolding reaction or by the addition of intermediates to each other. In this case the rate depends on the protein concentration.
raised to a power greater than one, indicating that intermolecular collisions control the rate (Wang, 1999; Wang, 2005).

The rates of the processes are controlled by the solvent environment which also controls the outcome of the aggregation process. Changing temperature increases the rates of unfolding and intermolecular collisions as well as enhancing hydrophobic interactions, which favors the formation of irreversible aggregates. Changes in pH can also alter the pathways by changing the stability of the native state ($K_i$) and the association rates (Jang, Hall, and Zhou, 2002b). Decreasing the charge on the protein increases the association rates due to decreasing intermolecular repulsion. Shaking and bubbling causes the formation of a large air-water interface which often adsorbs proteins. Air provides a hydrophobic surface which can cause the protein to unfold and expose its hydrophobic groups, leading to protein destabilization and the formation of the intermediate $I$. All these processes lead to enhanced protein aggregation. On the contrary, addition of salts or other additives often stabilizes the protein against unfolding and lowers the rates of aggregation (Wang, 1999; Wang, Nema, and Teagarden, 2010). We will investigate the possibility of mapping these changes in solution conditions to variable parameters in computer simulation experiments.

2.3 Simulation techniques for protein folding, misfolding, and aggregation

Initially, researchers focused on the ‘Protein Folding Problem’ (PFP), determining the pathways by which the protein folds naturally into its final native conformation. This work goes hand in hand with determining the native three-dimensional structure, or what is known as ‘Protein Structure Prediction’ (PSP), only from the primary structure. These two terms are interchangeable in current literature, usually meaning PSP (Lopes, 2008). Once the basic fundamentals of folding were understood, research was extended to aggregation mechanisms. However, characterizing the structural and dynamic properties of protein chains complexes and aggregates presented a challenging problem in structural biology (Dokholyan, 2006). The time of protein folding and aggregation depends on many factors including the size of the protein, its multidomain complexity, types of secondary structure upon folding, and more. It is however believed that folding and aggregation occur within
Chapter 2. Background

 microseconds (Kubelka, Hofrichter, and Eaton, 2004). Based on theoretical and experimental studies, it was suggested that the speed limit of proteins can be approximated to $N/100$ microsecond for a generic single domain protein with $N$ residues. This time-speed will gradually decrease with increasing size of proteins (Javidpour, 2012). Accordingly, proteins fold and aggregate in time and length scales inaccessible to conventional experiments. This makes computational modeling a needed complementary to experimental analysis.

Studying the protein folding problem using computational models is not without its challenges. The need of quantum mechanical features means that computational power is required in order to simulate even a small chain of amino acids for a nanosecond of real time. This computational cost is either very expensive or unreachable so far. Accordingly, an optimized sacrifice must be made between computational cost and accuracy level while modeling protein chain and protein systems, depending on the object of the study. We first present the biological search problem, i.e., the Levinthal’s paradox, in terms of computational challenges. Next we describe the ensemble where the simulation takes place. Protein folding and aggregation either occur at the tip of ribosomes, in the human body, in industrial reactors, in a lab beaker/flask, or in a medication ampule. Here we look at the sampling ensembles popular in computational studies. Computational experiments involved around proteins are hampered by computational cost, thus we will introduce advancements and proposed techniques to solve this problem by reducing the computational intensity, with special focus on replica-exchange method. In subsection 2.3.3 we introduce the main techniques for calculating the progression of the system. In subsection 2.3.5, we present the most popular protein models used ranging from all-atom to coarse-grained models. Subsection 2.3.6 will focus on the various interaction potentials employed in simulation experiments. Subsection 2.3.7 describes the solvent models implemented in simulation and finally subsection 2.3.8 outlines various forms of interaction potentials with focus on discontinuous molecular dynamics.

2.3.1 The search problem

The protein folding problem is effectively a search for the native state, which corresponds to the free energy minimum state over the internal degrees of freedom. A coarse-grained
Simulation techniques for protein folding, misfolding, and aggregation

parameter, ‘q’, also known as the order parameter is introduced as a measure of the ‘nativeness’ of the structures, in other words the similarity to the native state. The free energy of the proteins can be expressed as a function $F(q)$, as given by the following equation:

\[
F(q) = E(q) - TS(q)
\]  

In equation 2.1 $E(q)$ is the thermal averaged effective-energy function, $S$ represents the entropy, and $T$ represents the temperature. Averaged-out fast degrees of freedom are included in $E(q)$ (solvent motions for example). Using the remaining degrees of freedom $x_i$ (backbone dihedral angles for example), the energy function can be written explicitly such as:

\[
H = H(x_i)
\]

This is an energy minimization problem to find the low-energy structure which is slightly different from a canonical simulation. Practically, both $H(x_i)$ and $x_i$ represent several obstacles. Firstly, the potential function $H(x_i)$ is unknown. However, researchers have designed and introduced many empirically-based potential functions that proven to be useful in studying computational folding; Skolnick and coworkers (Skolnick, Kolinski, and Ortiz, 2000) derived protein specific pair potentials based on weak sequence fragment similarity. Their work showed an advancement over using the quasi-chemical pair potential by yielding more structures with a root-mean-square deviation less than four angstroms from the native state. Similarly, Chen and coworkers (Chen, Yang, and Shakhnovich, 2007) showed a significant improved ability in overcoming energetic barriers which has led to a lower energy native state with smaller root-mean-square deviation in native structures in both simulated annealing and replica exchange Monte Carlo (REMC) simulations. These knowledge-based set moves do not only improve the thermodynamics of the folding but also the kinetics by reaching the energy minima faster. On the other hand, Kolinski and coworkers (Kolinski and Skolnick, 1992; Kolinski, Godzik, and Skolnick, 1993; Skolnick, Kolinski, and Ortiz, 2000) managed to precompute and store energy contribution in their simulations, and save time by discretizing the conformational space by fixing the peptides on a lattice, or by setting a small range of the backbone dihedral angles derived from clustering dihedral
Chapter 2. Background

angles sampled from experimental structures (Chen, Yang, and Shakhnovich, 2007). Other designed potentials include Miyazawa-Jernigan potential (Miyazawa and Jernigan, 1985) and HP models which appear to be successful in studying protein folding and aggregation problems. Once the estimate of the potential energy function has been selected, the other tremendous problem presents itself; given the enormous number of degree of freedoms, the complexity of the problem greatly increases with increasing ‘ruggedness’ of the energy landscape. Non-native contacts can form along native-contacts in the folding pathway. This has been shown both experimentally and in simulation (Sosnick et al., 1994). Also, compact metastable structures known as molten-globules, are found to be characteristic intermediate in many folding processes, and cause the structures to be trapped in local energy minima. The problem is finally exacerbated by errors in the approximation of the energy function which makes solving equation 2.1 practically impossible without additional implementation and optimization.

2.3.2 Sampling ensembles

In this work, we consider simulations at constant temperature which correspond to either the canonical ensemble \((NVT)\) or the isothermal-isobaric ensemble \((NPT)\). The first ensemble occurs at constant number of particles, temperature and volume. The \(NPT\) ensemble on the other hand is at constant particle number, pressure and temperature. By solving the equations of motion as is done in molecular dynamics simulations, energy is conserved and simulations correspond to the microcanonical ensemble \((NVE)\). Thus, in order to keep the temperature constant the ensemble is coupled with a bath or thermostat (Scheraga, Khalili, and Liwo, 2007).

One of the methods employed to overcome the search problem is canonical sampling. Simulations can be performed in the canonical ensemble independently of the model used (refer to section 2.3.5) and the methods of sampling (Molecular dynamics or Monte Carlo shown in section 2.3.3). This approach aims at overcoming the rugged energy landscape by sampling the low-energy conformations according to a weighting function given by Boltzmann’s law (Liwo et al., 2008):
$P(X_i) = e^{-E(X_i)/k_BT}$

$P(X_i)$ being the probability of occurrence of the conformation described by the variables $X_i$, $E(X_i)$ being the energy of this conformation, $k_B$ is the Boltzmann’s constant, $T$ is the absolute temperature, so the Boltzmann factor is written as

$$W_B(x;T) = W_B(E;T) = e^{-\beta E}$$

with $\beta$ being the reciprocal of temperature equal to $1/k_BT$. The canonical probability distribution of potential energy $P_B(E,T)$ is given by the product of the density of states $n(E)$ and the Boltzmann weight factor $W_B(E,T)$ as

$$P_B(E;T) \sim n(E)W_B(E;T)$$

Two methods, the Monte Carlo (MC) method of Metropolis and Molecular Dynamics (MD), are used for sampling the canonical ensemble (more details about MD and MC are available in section 2.3.3). These algorithms are expected to yield $P_B(E;T)$ in theory.

Other generalized algorithms have been developed to overcome the energy barriers. Multicanonical algorithms or ensembles (MUCA) are a class of methods in which a logarithm for the density of states replaces the energy in the Metropolis criterion (Hao and Scheraga, 1998) of Monte Carlo algorithms such that energy levels are sampled with an equal probability (if the exact density of states is used) (Berg and Neuhaus, 1992). When all energy levels are sampled with the same frequency, sampling of rugged energy landscape can be improved. This method has proven to be effective for a wide range of applications in the protein folding pathways (Mitsutake, Sugita, and Okamoto, 2001) including all-atom (Mitsutake, Sugita, and Okamoto, 2003b) and simplified models (Nanias, Czaplewski, and Scheraga, 2006) in accordance with molecular dynamics and Monte Carlo.
2.3.3 Molecular Dynamics and Monte Carlo

The Monte Carlo (MC) method of Metropolis and the molecular dynamics (MD) are the two main techniques used in protein simulations (Javidpour, 2012).

Monte Carlo simulation consists of taking a random walk in the conformational space according to a specified weighting factor (Liwo et al., 2008). For the canonical ensemble, the method is summarized as follows: once the system is set-up, the initial conformation is perturbed at random, and the energy of the new conformation is calculated. If the energy of the new conformation is lower than the old one, the new conformation replaces the former; if not, the old conformation replaces the former with probability $\exp(-E/k_BT)$ where $E$ indicates the difference between the two conformations. This is an iterative process that goes on until the ensemble averages converge. The ensemble averages are updated every given number of steps. Due to the random sampling during energy evaluation steps, a lot of proposed moves are unaccepted depending on the maximum allowed displacement for each move. This hampers the efficiency of the approach. MC is ineffective at sampling rough energy landscapes as it is difficult to escape from configurations located in local energy minima. With time, optimization methods were developed to speed-up MC simulations, which have proven to be effective in studying various aspects of the protein folding problem.

Molecular Dynamics (MD) is an invaluable algorithm for studying protein folding and unfolding events and pathways, native state configuration, interactions underlying such events, and their time dependence due to the accuracy of the potential energy function. In this approach, numerical integration methods are used to obtain momenta and coordinates of the particles along their trajectories by means of solving Newton’s or Lagrange’s equations. In some cases where the solvent is not treated explicitly, Langevin’s equation is solved to account for the solvent by introducing a random force. The initial state of the system can be either unfolded proteins, folded chains in the native structure, or any random structure (Javidpour, 2012). Once the system is designed, the initial velocities are chosen from Maxwell-Boltzmann distribution at the simulation temperature.

The first application of MD in the field of proteins, specifically folded globular protein bovine pancreatic trypsin inhibitor, was conducted by Mccammon, Gelin, and Karplus (Mccammon, Gelin, and Karplus, 1977) in 1977. Since then, numerous MD simulations have
2.3. Simulation techniques for protein folding, misfolding, and aggregation

dealt with the protein folding problem, and various methods were applied in order to overcome the computational intensive requirements of MD simulations.

Solving the MD equations is computationally expensive and was only made possible to use by great advancements and development of computing power (Scheraga, Khalili, and Liwo, 2007). One shortcoming of MD is that the simulation might sample only a small region of the phase space, with a time-scale depending on the complexity of the potential used (Javidpour, 2012). Space-scale is not the only challenge to MD, rather the time scale of protein folding, association and aggregation vary from seconds to month, making them unreachable by traditional molecular dynamics simulations (Brooks et al., 1983; Pearlman et al., 1995). Using atomistic systems requires tremendous computational power, but improve greatly our understanding of protein folding. Other techniques emerged to provide a smarter choice of the simulation conditions by introducing certain simplifications (corase-graining for example) which increase the access to phase space region with shorter times. These constraints are on a conformational, energetic, and interaction levels. Combining these constraints with sampling methods introduced in section 2.3.4 in MC or MD simulations appears to be an efficient approach for overcoming these problems.

2.3.4 Other Sampling ensembles: replica-exchange technique

Many other algorithms have been developed to improve sampling systems described by rugged energy landscapes, these include umbrella-sampling methods (Torrie and Valleau, 1977) coupled with the weighted histogram analysis (WHAM), replica-exchange methods (REM) (Hansmann, 1997; Sugita and Okamoto, 1999; Sugita, Kitao, and Okamoto, 2000; Gront, Kolinski, and Hansmann, 2005) with multiplexing variant (MREM) (Rhee and Pande, 2003; Nanias, Czaplewski, and Scheraga, 2006). Other sampling methods such as locally enhanced sampling (LES) were specifically useful in ligand diffusion or loop optimization in proteins (Hornak and Simmerling, 2003), as well as ligand binding problems (Czerminski and Elber, 1991). These methods allow targeting a specific region of the conformational space while keeping the rest of the system intact (Liwo et al., 2008). Moreover, a multitude of ensembles are generated by combining existing ensembles with some new techniques which allows the overcoming of local energy minima. One of the most popular and useful techniques is replica-exchange (RE).
The replica exchange method, also referred to as multiple Markov chain method (Mitsu-
take and Okamoto, 2000) or parallel tempering method, is known as an effective sampling
method used to overcome the problem of rugged energy landscape and kinetic energy traps
(Hansmann, 1997; Sugita and Okamoto, 1999). Replica-exchange has been used to study
glassy systems such as biomolecules, spin glasses, protein folding, clusters, zeolite structure
solutions and many others (Kofke, 2002). The general idea of this sampling technique is
that replicas overcome energy barriers by performing random walks with respect to tem-
perature. Accordingly, this technique involves the simulation of \( n \) replicas at \( n \) different
temperatures in parallel (\( T_1, \ldots, T_n \)). As simulation proceeds, replicas located at nearby tem-
peratures (subscript 0 and 1) are occasionally exchanged, with a probability of swap accep-
tance

\[
P = \min[1, \exp(-\beta_0 - \beta_1)(U_0 - U_1)]
\]

(2.6)

where \( \beta_i \) is the reciprocal temperature \( 1/kT \) of system \( i \), \( k_B \) is Boltzmann’s constant, and \( U \)
is the potential energy of the system. In general, the replicas 1 and 2 may or may not be
temperature adjacent. It is necessary that non-adjacent exchange occur in order to escape
from metastable states (Fiore, 2011).

The exchange probability is determined by the fraction overlap between the corresponding
energy histograms, which increases with decreasing temperature difference. The efficiency
of this method depends strongly on the set of temperature chosen. If the difference between
the selected temperatures is very high, the high-temperature runs will sample configura-
tions unaccepted by the low-temperature runs; consequently the probability of acceptance
will decrease and the method will be ineffective (Kofke, 2002). With increasing system size,
more temperatures are required. This is because the average energy increases with the sys-
tem size while the width of the histogram is proportional to the square root of the system
size. This causes a decrease in the swap rate due to the narrowing of the energy distribu-
tions, as such more replicas (e.g., temperatures) are needed to simulate larger systems.

Since the probability of acceptance decreases with increasing temperature difference, it is
more probable that the exchange occurs between neighboring temperatures. This method
was found to be effective in overcoming energy-barriers at low-energy state and efficient in
sampling larger conformational space by allowing configurations not to get trapped in local energy minima (Kofke, 2002).

The key to optimizing temperature selection lies in balancing between large number of parallel processors needed to simulate temperatures and the probability of overlap for accepted swaps which should not be too low. Several studies focused on the optimization of temperature selection and the acceptance probability of replica-exchange.

Some relatively straightforward procedures suggest that the temperature set should follow a geometric or arithmetic progression. Such schemes require setting only the minimum and maximum temperature. In the ad hoc distribution starting from a certain temperature $T$, the temperature $T_1$ is calculated in such a way that the exchange probability between adjacent replicas is about 30% (Fiore, 2011). In most of these procedures, the progression used gives a rough estimate of the temperature set. Repeating the simulation for a short period of time allows the optimization of the set for a given swap acceptance rate. However, this is an iterative process as the swap rate depends on the equilibration of the low-temperature runs, which improves as temperatures are optimized. If the simulation runs for a short time it is not totally equilibrated, and once the temperature set are optimized, the swaps might change if the run goes for longer time.

In 2005, Predescu and coworkers (Predescu, Predescu, and Ciobanu, 2005) defined the effective fraction as the probability of the configuration from the lowest index reaching the highest index in a simulation. On the other hand, many researchers tried to decide the optimum percentage for the average accepted swaps. Predescu and coworkers (Predescu, Predescu, and Ciobanu, 2005) suggested a value roughly around 40% (specifically 38.74%). Although this seems a little too exact to apply to all simulations, Predescu and coworkers argue that this remains roughly the optimal probability for most systems with continuous distributions that are likely to be encountered in practice. Kofke (Kofke, 2002) correlated the average swap acceptance probability of replicas with its entropy difference. Predescu et al. argue that for a constant heat capacity, the temperature-set should follow a geometric progression, thus they relate the swap acceptance probability to the heat capacity. Later-on, Kone and Kofke (Kone and Kofke, 2005) argued that the temperatures in the replica-exchange method should swap in a 20% probability for a constant heat capacity. Katzgraber et al. (Katzgraber et al., 2006) proposed a feedback-optimized parallel tempering Monte Carlo that generates the set of
temperature by measuring the replica diffusion through the temperature space, this work is the extension of the method introduced by Predescu et al. in 2005. In 2008, Sabo et al. (Sabo et al., 2008) introduced a new method, the constant entropy method (CEM), which does not only exhibit the same advantages of the feedback parallel method, but also seems to be less computationally intensive.

Two modified method of RE, referred to as the multiplexing variant of Replica-exchange method (MREM), are known to enhance sampling (Rhee and Pande, 2003; Nanias, Czaplewski, and Scheraga, 2006). Instead of temperature-swap, restraints are exchanged between replicas in the first method (Gront, Kolinski, and Hansmann, 2005; Sugita, Kitao, and Okamoto, 2000), and Hamiltonian-exchange occurs in the second modified method (Sugita, Kitao, and Okamoto, 2000; Liu et al., 2006).

Other methods that combine multicanonical algorithms (MUCA) with replica exchange (RE) are also employed. In one method, the replica exchange multicanonical algorithm (RE-MUCA), replica exchange generates the initial density of states as opposed to estimating it. Once the density of states is known all ensemble-averages can be computed (Liwo et al., 2008). In another, a combination of replica exchange and multicanonical algorithm is used to calculate the initial density of states. This is known as the multicanonical algorithm replica-exchange method (MUCAREM). Replica exchange simulations are performed about various regions in energy space with small replicas respectively (Mitsutake, Sugita, and Okamoto, 2003b). REMUCA and MUCAREM were implemented in various studies, but the feedback was not consistent. Mitsutake and coworkers (Mitsutake, Sugita, and Okamoto, 2003a) reported 4 times reduction in computational cost between REMUCA and MUCAREM for a 5 ns simulation time of a complex system of 17-residue helical peptide. On the other hand, Liwo and coworkers (Scheraga, Khalili, and Liwo, 2007) implemented MUCA, REMUCA, and MUCAREM algorithm in their coarse-grained UNRES forcefield, and found that they do not offer significant advantage over the multiplexing variant of replica exchange method (Nanias, Czaplewski, and Scheraga, 2006). Rather, implementing those methods appeared to be computationally expensive, and the calculated density of states was unstable in low-energy regions.
2.3.5 Protein Models: from all atoms to coarse-graining

An alternative solution to the search problem is a structural simplification of the protein model, or coarse-graining. A variety of protein models have been used and these can be divided into three main categories: the most realistic models are complex high-resolution models (Hall et al., 2006) or all-atom models as they take into account almost every atom in the system. This type of models has provided a wealth of information on the process of protein unfolding, structural properties on the non-native ensembles, and protein-protein interactions (Hall et al., 2006). The second class of representation are simplified coarse-grained models in which a single sphere or bead represents a single amino acid. Despite their simplicity, these low-resolution models have allowed valuable insight into a variety of interesting phenomena including structures of intermediates in the folding pathways, such as the molten globules (Hammarstrom et al., 1999), and structural low-energy conformations in the native-like folded ensembles (Dokholyan, 2006). The third class of models presents an intermediate resolution between one bead per amino acid and all-atom representations (Hall et al., 2006). Intermediate-resolution models can take many forms and will be discussed in section 2.3.5.2.

Simplified models offer a resolution to time- and length-scale biological phenomena, i.e. large-scale dynamics, and simulation of long-time mechanisms (Abkevich, Gutin, and Shakhnovich, 1995; Skolnick, Kolinski, and Ortiz, 1997; Zhou and Karplus, 1997; Ding and Dokholyan, 2005; Dokholyan et al., 2003; Smith, Hall, and Freeman, 1997). The interaction potential between the beads is derived empirically from experimental work (Dixon et al., 2004), physics (Miller et al., 1992; Srinivasan and Rose, 1999), or from protein structure data-bank (Go and Abe, 1981). These will be detailed in section 2.3.5.3. The size of the system addressable with computer simulations has gradually increased due to the large advancements made in computer power. Early simulations of protein folding necessarily relied on minimal models of the protein chain. These simplified models coarse-grained the level of representation, where a monomer can sample a part of a residue, a single residue, several residues, or even a complete peptide. With increasing computational power, and the increase interest in the subject, all-atom models start emerging where each atom of the chain is individually represented, including the representation of the solvent in an explicit or implicit way. The choice of the model relies on the objective of study, whether it aims to simulate a real protein such as
Chapter 2. Background

the bovin pancreatic trypsin inhibitor (BPTI), or to simulate the generic behavior of proteins in a solution. The level of coarse-graining can be divided into three: high resolution models, intermediate resolution models, and low resolution models, depending on the level of complexity and representation.

2.3.5.1 High resolution models

The early all-atom simulations were performed in the 1990s. Daggett and Levitt (Daggett and Levitt, 1992; Daggett and Levitt, 1993) simulated the bovine pancreatic trypsin inhibitor in 1992 at 423 K for 550 picoseconds in its oxidized form and at 498 K for 284 picoseconds in the reduced form. At higher temperatures, while the $\alpha$-helix lost some of its conformation, the $\beta$-sheets gained some secondary structure and exhibited higher stability. In 1992, two unfolding pathways towards the denatured state of BPTI were detected. In one pathway, first $\alpha$-helices unfolded followed by $\beta$-sheets. This was simulated at neutral pH. In the second pathway the order of unfolding was reversed at low pH. The unfolding of the Hen Egg White Lysozyme appeared in 1992 with a simulation done at 500 K for about 200 picoseconds (Mark and Vangunsteren, 1992). The simulation showed an initial expansion of the protein associated with an increase in solvent accessible surface area, coupled with a sharp decrease in helical hydrogen bonding. This study was revisited in 1995 for a detailed analysis of the unfolding pathway and characteristics (Hunenberger, Mark, and Vangunsteren, 1995). The denaturation of myoglobin was also simulated with runs at 358 k (Tiradorives and Jorgensen, 1993), along with the functionality of its native state (Makarov et al., 2000). At 298 K the protein lost its crystal structure due to the removal of the heme group causing the loss of helical structures from 85% to 55% (Griko et al., 1988). Daggett and coworkers (Daggett et al., 1996) explored the atomic details of chymotrypsin inhibitor 2 (CI2) unfolding with extreme temperature simulations (Li and Daggett, 1994; Li and Daggett, 1996; De Jong et al., 2002) using all-atom models. In 2002, a general unfolding pathways for CI2 was derived independently from temperature. Day et al. (Day et al., 2002) showed that CI2 unfolds from the native state to the transition state by a disruption of the core packing, followed by a breathing motion of the N-terminals. The transition state is characterized by a non-stable hydrophobic core and a loss in some of its secondary structure. At higher temperatures, the active site unfolded and the core became completely disrupted. Protein A (Boczko and
Brooks, 1995; Guo, Brooks, and Boczek, 1997; Alonso and Daggett, 2000), ubiquitin (Alonso and Daggett, 1998), Protein G (Sheinerman and Brooks, 1998), SH3 Domain (Tsai, Levitt, and Baker, 1999; Gsponer and Caflisch, 2001) were also investigated by all-atom molecular dynamics simulations as a complement to experimental studies.

Although these successful studies presented the most realistic representation of proteins under investigation, their complexity makes them computationally intensive. Therefore, all the folding and aggregation problems that occur on long time scales or large conformational changes are inaccessible to such detailed models (Hall et al., 2006). An example is the simulation performed by Duan and Kollman (Duan and Kollman, 1998). Two full months where needed to perform a 1-microsecond simulation of the folding of a 36-residue protein in aqueous solution with more than 3000 water molecules, going from a structure with some native turns to a partially folded intermediate, using 256 processors.

### 2.3.5.2 Intermediate-resolution models

Intermediate-resolution representation of proteins is a good compromise between the highly complex all-atom models, and the simplified low-resolution homo- and heteropolymer chain models (Smith and Hall, 2001). Many intermediate-resolution models have been developed by increasing the number of beads used to represent the protein structure. The models were built using three-beads (Liwo et al., 1993a; Liwo et al., 1993b; Liwo et al., 1997c; Liwo et al., 1997a; Liwo et al., 1997b; Liwo et al., 1998; Hardin, Luthey-Schulten, and Wolynes, 1999; Hardin et al., 2002; Wallqvist and Ullner, 1994), four-bead and five-beads (Takada, Luthey-Schulten, and Wolynes, 1999; Sun et al., 1992; Sun, 1993; Sun, Thomas, and Dill, 1995; Sun, 1995; Irback, Sjunnesson, and Wallin, 2000; Irback, Sjunnesson, and Wallin, 2001; Favrin, Irback, and Wallin, 2002). Kolinski et al. (Kolinski and Skolnick, 1992; Kolinski, Godzik, and Skolnick, 1993) model each residue in the chain as two beads, with one side-chain bead and one backbone bead. Another intermediate-resolution model is the three-bead model introduced by Wallqvist and Ullner in 1994 (Wallqvist and Ullner, 1994), with one interacting site, the backbone bead, and one or two side-chain beads depending on the size and complexity of the residue. Using MC simulations, two small proteins, avian pancreatic polypeptide and a parathyroid hormone-related protein were simulated from the high temperature extended
state. The results obtained are consistent with experimental results and reproduced satisfactorily the native structures. Smith and Hall argue that the four-bead model (three backbone beads and one side-chain) put forward a more accurate representation of the protein backbone structure, compared to the lower-resolution models which allows independent interactions between side chain and backbones (Smith and Hall, 2001; Sun, 1993; Takada, Luthey-Schulten, and Wolynes, 1999).

The four-bead model presented by Smith and Hall (Smith and Hall, 2001) has an $N$, $C^{\alpha}$, $C$, and $R$ beads representing the amino acid’s amide nitrogen and hydrogen, the $\alpha$-carbon and its hydrogen, the carbonyl-carbon and oxygen, and the side chain respectively. This three-bead backbone with single-bead side-chain model is used in discontinuous molecular dynamics simulations to study the competition between folding and aggregation in multi-chain system with eight 16-monomer chains per system. Two different folding pathways were equally favored by the chain: one through a trimeric intermediate and the other involving the association of two dimers. However, for the four-peptide system, the chain heavily favored one pathway over the other. The difference in the folding tendencies suggested that the presence of other proteins in a system has a strong effect on the folding properties and dominant pathways. Moreover, spontaneous fibril formation was investigated in polyalanine-based peptides of the sequence $Ac-KA_{14}K-NH_2$ (Smith and Hall, 2001). This is a 16-mer peptide composed of 14 alanine (A), with lysines at both termini. The folding behavior of a single peptide exhibits generic features of the real polyalanine chain. This includes the formation of $\alpha$-helices, $\beta$-strands, and unfolded coils depending on the solvent conditions (Nguyen, Marchut, and Hall, 2004). First, the interaction strength $IR$ was defined as the ratio of strength of the side-chain’s hydrophobic interactions and backbone hydrogen bonding interactions, $IR = \epsilon_{HP}/\epsilon_{HB}$. In effect $IR$ measures solvent conditions, with $IR = 0$ corresponding to the vacuum. The effect of $IR$ on the folding transition was monitored in order to choose the optimum parameters for multiple chain system. A value of $IR$ equal 1/10 was chosen for systems containing 12, 24, 48, and 96 polyalanine chains. The effect of temperature and concentration on the formation of $\beta$-helices, $\beta$-sheets, amorphous aggregates, and fibrils was also investigated (Nguyen and Hall, 2004). Simulation results showed that the percentage of $\alpha$-helix increases with a slight increase in temperature, and starts decreasing towards zero at higher temperatures after going through a maximum. The
dependence on peptide concentration showed a similar behavior. For concentration greater than 0.5 mM, the helices did not form. For lower concentrations, α-helix formation was maximized around the temperature $T^* = 0.09$. On the other hand, the formation of fibrils occurred at high temperatures and concentrations, with the critical temperature of formation decreasing with increasing concentrations of peptides. These results are consistent with experimental data (Antzutkin et al., 2000; Antzutkin et al., 2002; Harper, Lieber, and Lansbury, 1997). In the same work, Hall et al. studied the kinetics of fibril formation and showed that it is nucleation dependent, which is also in agreement with experiments (Jarrett and Lansbury, 1993). Finally, a phase diagram in the temperature-concentration plane was mapped out to delineating regions where random coils, α-helices, β-sheets, fibrils, and amorphous aggregates are stable. The diagram can be used as a benchmark for experimentalists either for studying fibrilization events, or avoiding it.

### 2.3.5.3 Low-resolution models

The third level of coarse-graining is simplified low-resolution models. In these simplified models one bead represents a single amino acid (including its side chain). Therefore the peptide chain is simply a series of connected beads. Moreover, two-bead resolution models can be considered as a part of the intermediate-resolution models class as shown in the previous section, or low-resolution model class. This structural discretization provides an alternative solution to the search problem and offers an extension of the timescale of simulation compared with intermediate- and high-resolution models (Kolinski and Skolnick, 2004; Nielsen et al., 2004). This is in fact due to the reduction in the degrees of freedom. Although much is left out or estimated by even the most complex minimal models, low-resolution models greatly enhanced our understanding of protein folding (Onuchic, LutheySchulten, and Wolynes, 1997). If the aim of the study is to mimic protein generic behavior in order to develop a good understanding of the process, coarse-grained models with simplified representation of protein chains allow the extraction of general principles regarding the thermodynamics and kinetics of aggregation. The varying breadth in time and length scales involved in protein mechanisms lends to the use of models with different levels of resolution, with each level aiming to probe a different element of the process.
The work on simplified one-bead (Wilson and Doniach, 1989; Honeycutt and Thirumalai, 1992; Guo, Brooks, and Boczko, 1997; Guo and Thirumalai, 1995; Guo and Thirumalai, 1996; Guo and Thirumalai, 1997; Veitshans, Klimov, and Thirumalai, 1997; Klimov, Betancourt, and Thirumalai, 1998) and two-bead models per amino-acid (Levitt and Warshel, 1975; Miyazawa and Jernigan, 1982; Rey and Skolnick, 1991) were pioneered by Levitt, Warshel, Kuntz, Crippen, Kollman, Scheraga and coworkers, and Go and coworkers (Go, 1983). The bovine pancreatic trypsin inhibitor, extensively studied in experiments and simulated by all-atom simulations (as shown in section 2.3.5.1), was revisited by Levitt and Warshel using two-bead model interacting via empirical potentials derived by correlations from data on dipeptide structures in off-lattice simulations (Levitt and Warshel, 1975). The type of interaction of protein-models is discussed in details in section 2.3.6. This work was initially criticized because the interaction was biased towards the formation of the native state, nonetheless it showed for the first time that folding can be represented using reduced-coordinate models (Hall et al., 2006).

Coarse-grained low-resolution models are divided into two types: on-lattice models where beads are confined to a lattice, and off-lattice models where proteins are allowed to move freely in the simulation box. Off-lattice models were not very popular in studying aggregation (Hall et al., 2006). Most of the low-resolution models used to study aggregation and folding are on-lattice models. This gave a motivation to put forward in this project an off-lattice coarse-grained model to study folding behavior and aggregation.

In off-lattice models, proteins are represented by a chain of beads or group of beads moving freely through continuous space. Jang and coworkers (Jang, Hall, and Zhou, 2004a; Jang, Hall, and Zhou, 2004b; Jang, Hall, and Zhou, 2002a) developed 3 off-lattice 2-bead models. Each model constituted of three antiparallel $\beta$-strands: the $\beta$-sheet, the $\beta$-clip, and the $\beta$-twist. Motivated by the formation of $\beta$-strand and the conversion of $\alpha$-helices into $\beta$-sheets during aggregation, they studied the thermodynamics (Jang, Hall, and Zhou, 2004b) and the kinetics (Jang, Hall, and Zhou, 2004a) of the assembly of four of the model $\beta$-sheet peptides. They introduced a 2-dimensional temperature-phase diagram for the tetrameric $\beta$-sheet complex (Jang, Hall, and Zhou, 2002a; Hall et al., 2006; Hall and Waggner, 2006). On another note, Ding et al. (Ding et al., 2002) modeled a system of eight Src-SH3 domain proteins, where each peptide chain is represented by an off-lattice two-bead model interaction
via the Go potential, with one bead representing the backbone, and the other the side chain. The aggregated chains formed a fibrillar double β-sheet structure with inter-β-strand spacing distances and inter-β-sheet distances equal to the spacing found in experiments. This shows that off-lattice models provide insight to the aggregation process with more details than lattice models (Hall et al., 2006). However the use of Go-potential with biasing towards forming the native conformation is still criticized. Alternative interaction functions are introduced in section 2.3.6 with more realistic approaches. In 1998, Hall and Gupta (Gupta, Hall, and Voegler, 1998) presented their work on the competition between refolding and aggregation. The protein is simulated as a chain of hydrophobic (H) and polar (P) beads arranged in a self-avoiding random walk, fixed on a two-dimensional square-lattice model. Monte Carlo simulation of a system of 40 heteropolymer chain showed that aggregation arises from association of partially-unfolded rather than totally denatured chains, as it was accepted in the early years.

Another technique emerged for coarse-grained models which simplifies the system even further by confining the chain of beads to a lattice, with beads interacting by one of two types of interaction, either attraction or repulsion. This extended simplification which induces structural restriction in the degrees of freedom was innovated by the models of Go and coworkers (Go and Taketomi, 1978; Go and Taketomi, 1979b; Go and Taketomi, 1979a). They took the additional step of restricting the beads to points on a two-dimensional lattice and used only two types of beads. This is the simplest model of protein structure, referred to as coarse-grained on-lattice model. The lattice could be in 2-D plane or 3-D space. Square and cubic lattices are the most commonly used (Javidpour, 2012), but sometimes other forms of lattices are implemented including triangular (Li, Zhang, and Chen, 2005) and hexagonal (Jiang and Zhu, 2005). In the lattice, adjacent amino-acids must occupy adjacent position on the lattice, with a single position filled by one amino acid only. Confining the chains to a lattice discretizes the conformational space and pose strong constraints on the angles between the covalent bonds, hence providing another solution to the search problem (Dokholyan, 2006) by reducing the conformational space to be sampled. This also allows the exact enumeration of possible conformations including the lowest energy structure. But at the same time, it might not capture all the characteristic behavior of proteins folding, misfolding, and aggregation.
Simulations using on-lattice models were pioneered by Go et al. in the late 1970 (Hall et al., 2006) with 2-D and 3-D lattice models. In the 1980’s, kolinski et al. identified the minimal set of interactions required to fold a protein-model into native-like motif structure. Using Monte Carlo simulations and three-dimensional lattice models, they showed that peptide chains can fold without assigning tertiary interactions that favor the native conformation (Kolinski, Skolnick, and Yaris, 1986; Kolinski, Skolnick, and Yaris, 1987; Skolnick, Kolinski, and Yaris, 1988; Skolnick, Kolinski, and Yaris, 1989b; Skolnick, Kolinski, and Yaris, 1989a). Harrison et al. (Harrison et al., 1999; Harrison et al., 2001) studied the thermodynamic association with the aggregation of pairs of two- and three-dimensional HP chains by exact enumeration of all possible conformations. They found that a single mutation in the protein sequence might induce aggregation. Also, monomers with high stability might increase the aggregation propensity by forming homo-dimers, a known feature of prion aggregation. Broglia et al. (Broglia et al., 1998) investigated the aggregation process using two identical on-lattice 36 amino acid chains interacting via the MJ potential. They showed that partially folded intermediates with strong residue-residue bonds drive the aggregation process. Istrail and coworkers (Istrail, Schwartz, and King, 1999) studied aggregation in a two-chain system with a modified interaction, which takes into account repulsive interaction between solvent sites and hydrophobic residues. They showed that the ratio of residues with repulsive interaction over the residues with attractive interactions affects the aggregation propensity. For a small ratio, chains are more susceptible to aggregate. Dima and Thirumalai (Dima and Thirumalai, 2002) studied the aggregation of monomeric compact state to an oligomeric $\beta$-sheet state in a two-peptide chain system fixed on a 3-D lattice and interacting via HP potential. They reported a temperature-concentration phase diagram with 3 distinct ordered structures, with only one of them as the native conformation. They also argued that the aggregated structures formed directly from the denatured structures rather than the intermediate states. All previous simulations related to the aggregation process consist of multichain systems with very few peptide-models, up to 2 chains per system. Many groups put forward a more realistic multichain system by increasing the number of beads in a peptide on one hand (up to 20 residues), and increasing the number of chains in the system (up to 40 chains). Using a multichain system with twenty 8-monomer peptides, Combe and Frenkel (Combe and Frenkel, 2003) derived a protein phase diagram with 3 distinct phases.
A metastable gas-liquid transition was reported because the liquid-vapor transition lies below freezing temperatures. Gupta and coworkers (Gupta, Hall, and Voegler, 1998) simulated multichain systems containing 20 and 40 peptides with 20 monomers interacting via the HP potential. They reported that aggregation is induced by partially folded chains rather than totally denatured ones. Nguyen and Hall (Nguyen and Hall, 2002) used the same model in the previous study by Gupta and coworkers (Gupta, Hall, and Voegler, 1998) but with multichain systems up to 50 peptides at various values of packing fraction. The work was focused on the effect of the rate of chemical or thermal renaturation on the folding and aggregation behavior. They showed that the slow-but-finite cooling and infinitely slow-cooling methods give higher refolding yield than the quenching method, with the slow-but-finite cooling method being the fastest. Their predictions appeared to be consistent with experimental results on the lysozyme system (Yoshii et al., 2000; Hevehan and Clark, 1997). Istrail et al. (Istrail, Schwartz, and King, 1999) studied the aggregation propensity in an HP, 2-D on-lattice model, and argued that aggregability is primarily a property of kinetic intermediates, rather than native conformations, consistent with the experimental results on aggregation. A more complex representation of residue-residue interactions was implemented by Broglia et al. (Broglia et al., 1998) through the simulation of a three-dimensional lattice-model with residues interacting via the MJ potential, to study the role of partially unfolded chains in increasing the propensity of aggregation. Moreover, Harrison et al. (Harrison et al., 1999) investigated the thermodynamics of conformational change associated with aggregation of proteins. The model was used to simulate prion proteins as HP model-chains fixed on 2- and 3-dimensional lattices. Other two-dimensional and three-dimensional HP models have shown a behavior consistent with that of ‘real proteins’ (Dill et al., 1995; Dinner et al., 2000; Dobson and Karplus, 1999; Penas et al., 2015). Off-lattice models present undoubtedly a more realistic representation of proteins. Although allowing the chain to move freely in the conformational space makes it more computationally expensive, the off-lattice representation is better in capturing the generic behavior of the folding pathways and aggregation mechanisms.

Coarse-grained models are ideally suited for qualitative studies of protein aggregation due to the increase in computational costs associated with the mutichain systems (Bratko and Blanch, 2001). Also, the coarse-grained simulation have provided a conceptual solution to
the Levinthal’s paradox as the chains do not need to search through the enormous number of possible conformations and configurations to reach the native structure (Dinner et al., 2000).

### 2.3.6 Interaction potential

In this section, we present the interaction models used to describe protein simulation. In reality, proteins interact through a variety of complex interactions including electrostatic interactions, hydrophobic interactions, hydrogen-bonding, weak forces such as Van-der-Waals forces, and many other type of interactions. Mapping these interactions into a simulated model is not an easy straight-forward task. Many interaction potential have been derived with various approaches and different level of accuracy in representation. Obviously, different level of coarse-graining would require different types of interaction potentials. What follows is a review of the types of interaction potentials subdivided between all-atom simulation, intermediate-resolution models, and simplified models.

#### 2.3.6.1 High resolution Interaction potential

All-atom simulations are the most realistic representation of ‘real’ proteins (Javidpour, 2012). Accurate representation of the forcefield is needed to fulfill its purpose in studying even the most complex of dynamics. A small extension on the review of molecular dynamics (from section 2.3.3) is introduced here with the purpose of explaining the idea of forcefields and interaction potentials. As mentioned previously, molecular dynamics apply Newton’s equation of motion while calculating the trajectories of the system. For a system with \( N \) particles, the equation of motion of a particle \( i \) (or a bead) with mass \( m_i \) is equal to:

\[
m_i \ddot{\mathbf{r}}_i = \mathbf{F}_i \quad \text{with} \quad i = 1, 2, 3, \ldots, N \quad \ddot{\mathbf{r}}_i = \frac{d^2 \mathbf{r}_i}{dt^2} \tag{2.7}
\]

with,

\[
\mathbf{r}_i = (x_i, y_i, z_i) \tag{2.8}
\]
where \( r_i \) is the vector of Cartesian coordinates of the \( i^{th} \) atom, \( \ddot{r}_i \) the acceleration, \( F_i \) is the vector of forces acting on the \( i^{th} \) atom and \( N \) the number of atoms.

Newton’s equations of motion are most applicable in simple systems. For more complex systems, generalized Lagrange equations of motions are more efficient (Smit, Desmedt, and Frenkel, 1989). We introduce this equation to point out that when all-atom simulations are employed the forces are only potential forces (Scheraga, Khalili, and Liwo, 2007). A different equation describes what is known as Langevin or Brownian dynamics. This takes into consideration other type of forces (such as collision and friction forces) and is included to mimic the collisions of the solute molecule with its environment. Such forces are required to take into account the effect of the solvent, which is treated implicitly in the model. In fact whenever coarse-graining is employed, whether in MD or MC simulations with implicit solvent, the interaction energies are free energies of interaction determined by averaging over solvent degrees of freedom, this averaging process leads to a free energy. Equation 2.7 is reduced to equation 2.9 with the forces being equal to \( F_i \).

\[
F_i = -\nabla_{r_i} U(r_1, r_2, ..., r_N) - m_i \gamma_i \dot{r}_i + R_i(t) \quad \text{with} \quad i = 1, 2, 3, ..., N \tag{2.9}
\]

where, \( \dot{r}_i \) is the acceleration coefficient of atom \( i \), \( \gamma_i \) is the friction coefficient of atom \( i \), \( U \) is the potential energy of the system, \( -\nabla_{r_i} U \) refer to the potential force acting on atom \( i \), and \( R_i(t) \) is the vector of random forces due to the collision of atom \( i \) with the molecules of the solvent (not considered explicitly).

We next shown how the potential force in equation 2.9 derived. For all-atom models we use the most accurate potentials. Such forcefields are obtained using quantum mechanics calculations, and then benchmarked to experimental measurements of biomolecules structures (Javidpour, 2012). Many empirically derived all-atom potential functions including CHARMM (Brooks et al., 1983) (Chemistry at Harvard Molecular Mechanics), AMBER (Parelmnan et al., 1995) (assisted model building with energy refinement), GROMOS (Berendsen, Vanderspoel, and Vandrunen, 1995) (Groningen molecular simulation), and CVFF (Ewig, Thacher, and Hagler, 1999) (consistent valence force field). These molecular dynamics packages treat the solvent either implicitly or explicitly. The modeling of the solvent is discussed in section 2.3.7.
Like all parameters, the choice of the forcefield form is an optimization between accuracy and computational cost (Scheraga, Khalili, and Liwo, 2007). An approximation can be made for electrostatic-interactions, by neglecting the effect of polarization and electron-charge density. Other approximations related to bond angles are also implemented. However, care must be taken when implementing harmonic functions for bond-stretching and bond-angle bending because unreasonably large distortions of the bond angles might occur (Roterman et al., 1989). This might lead to a non-consistency in properties simulated with different forcefields (Yeh and Hummer, 2002; Sorin and Pande, 2005). The inaccuracy in forcefield inherited per-residue is propagated and multiplied by the amount of residues in the system causing a large deviation in the error. Therefore, the forcefield optimized in simplified models needs to be performed for all-atom forcefields (Fain and Levitt, 2003; Schug and Wenzel, 2006).

It is obvious that the protein is not alone in the system, rather it exist in an equilibrium with a solvent. Interaction between peptide-chains and the solvent are also very relevant to the protein dynamics and the intramolecular interactions.

2.3.6.2 Intermediate- and low-resolution interaction potentials

Many energy functions with different types of interaction have been used in simplified coarse-grained models. There are two major categories of potential models: discontinuous potential (DP) and continuous potentials (CP). The details of these potential models are discussed in section 2.3.8. On the other hand, the types of energy functions used in coarse-grained simulations can be loosely divided into four categories.

The first type of interaction models are general simplified empirical potentials that are meant to capture only simple interactions such as hydrophobicity, polar, and charged interactions. They are a function of the relative hydrophobicity of the side chains in intermediate resolution models (including 2-bead simplified models), or the amino acids themselves in simplified low-resolution models. The simplest type of generalized models are the Hydrophobic-Polar potential model, collectively known as the HP model. The second category includes potentials derived from protein structures. These are Go- or Go-hybrid potentials with a biasing towards the native structure, or elements from the native structure (Go and Abe, 1981;
2.3. Simulation techniques for protein folding, misfolding, and aggregation

Brown and Head-Gordon, 2004). The third class of interaction potentials are knowledge-based models derived from experimental strategies (Dixon et al., 2004). The final class of interaction potentials includes potentials derived from physics or physics-based models (Miller et al., 1992; Srinivasan and Rose, 1999). Usually the type of interaction in categories 1, 2, and 3 is discontinuous, while continuous potential is commonly used in category 4. Next we explore each interaction potential model in more details.

The simplest type of interaction potential in the first category is the Hydrophobic Polar (HP) model where there is only two types of potential: either attraction or repulsion. Beads are labeled with one of the two labels, ‘H’ standing for hydrophobic beads, or ‘P’ referring to hydrophilic or polar beads (attraction).

As mentioned in section 2.3.5.3 low-resolution models are of two types, lattice and off-lattice models. The HP interaction potential models are very popular when model-peptides are fixed on a lattice (Scheraga, Khalili, and Liwo, 2007). Giugliarelli et al. (Giugliarelli et al., 2000) investigated the influence of inter-residue interaction strength on the formation of native structures, non-native aggregates, and aggregates with native contacts using multiple two-dimensional lattice-peptides interacting via HP potential. They showed that once the interaction potential is set to give protein-like hydrophobicity, native states and non-native aggregates form equally in large amounts. Here, we refer to non-native aggregates as aggregates with contacts that are not similar to those found in the native state. This is also known as ‘prion-like’ aggregates. Other Monte Carlo simulations are performed by Patro and Przybycien (Patro and Przybycien, 1994; Patro and Przybycien, 1996) for 2-dimensional peptide chains fixed on hexagonal lattices. The peptide models have side chains with relative attraction and repulsion. They tested the effect of protein surface characteristics and protein-protein interaction energies on the structure of the aggregates formed. The Monte Carlo simulations by Istrail and coworkers (Istrail, Schwartz, and King, 1999) reviewed in subsection 2.3.5.3 employed 2-D lattice peptides interacting via the HP model. Here, the regular HP potential was modified by adding a repulsive interaction between solvent sites and hydrophobic residues to take into account the excluded volume effect. For other simulations using HP interaction potential, the reader can refer to section 2.3.5.3 for the work of Harrison et al. (Harrison et al., 2001), Patro and and Przybycien (Patro and Przybycien, 1994), Dima and Thirumalai, Combe and Frenkel (Combe and Frenkel, 2003), Gupta and
Hall (Gupta, Hall, and Voegler, 1999), and Nguyen and Hall (Nguyen and Hall, 2002).

The second class of interaction-potential models implemented in coarse-grained simulations are potentials derived from protein structures with biasing towards the native state or some of its characteristics. These include Go-type potential and its hybrids. Go potentials use the knowledge of the protein structure, where the native contacts (contact formed in the biological active state) are known. These contacts, along with some special interactions identified \textit{a priori} from the native structure, are represented by attractive potentials. The potential between other residues-pairs is repulsive (Cieplak, Hoang, and Robbins, 2002; Taketomi, Ueda, and Go, 1975; Hoang and Cieplak, 2000). Although this type of model greatly reduces the computational intensity with the interaction discretization and restrictions, strong discussions about the results obtained from using the Go potentials have been raised. The work by Taketomi \textit{et al.} (Taketomi, Ueda, and Go, 1975; Ueda, Taketomi, and Go, 1978; Go and Taketomi, 1978; Go and Taketomi, 1979b) on fibrils and amorphous aggregates discussed in section 2.3.5.3 employed peptide model interacting via a hybrid Go-type potential. Depending on the value of the bias parameter measuring the difference in strength between non-native and native contacts, the potential took various form: either square-well (SW), square-shoulder, or shoulder well. These forms of potential are detailed in section 2.3.8. In another study Bratko and Blanch (Bratko and Blanch, 2001) simulated multichain systems of six 27-monomer peptides fixed on a lattice and interacting via the Go potential. The global minimum free energy was identified for a dimers in the native state. Their work gave similar results to those found by Giugliarelli \textit{et al.} (Giugliarelli \textit{et al.}, 2000): non-native aggregates of three or more chains (what Giugliarelli \textit{et al.} refer to as ‘prion-like’ aggregate) exhibit high stability compared to aggregates with native contacts. Another hybrid Go-type model was introduced by Sorenson and Head-Gordon (Sorensen and Head-Gordon, 2002) with knowledge-based rotation about the $C^\alpha$ ... $C^\alpha$ virtual bonds. This model with biasing toward native secondary structure was used to study the kinetics of ubiquitin-like sequences (Sorensen and Head-Gordon, 2002), and L and G proteins (He and Scheraga, 1998). The Go models have been applied in a variety of aspects of the protein folding problem including the kinetics of folding and sequencing events (Hoang and Cieplak, 2000; Cieplak, Hoang, and Robbins, 2002), thermal unfolding of proteins (Cieplak and Sulkowska, 2005), and unfolding due to mechanical disturbance in force clamp (Luccioli \textit{et al.}, 2010).
The third type of potential are model-potentials derived from experimental strategies found in the protein structural databases. One example of these databases is the Protein Data Bank (PDB) where the native structure of many proteins at atomic resolution is found from x-ray crystallography and nuclear magnetic resonance (Berman et al., 2000).

The knowledge-based potentials are derived from protein structural databases, both in their form and parameterization (Kolinski and Skolnick, 2004; Sippl, 1993; Meller and Elber, 2001; Crippen, 1996). A large part of these is based on the Boltzmann principle to derive the components from the relevant distribution and correlation functions calculated from the Protein Data Bank (Sippl, 1993). The other approach is the derivation of the potentials for threading; the principle is to locate the native-like structures as the lowest in energy among a large number of decoys derived from the PDB. The potentials based on the Boltzmann principle also require calibration on known protein structures to obtain an appropriate balance of different energy terms. One example of these threading protocols is the linear programming optimization implemented by Meller and Elber (Meller and Elber, 2001) with double statistical filter for protein. Initially, this type of knowledge-based potential was used for predicting protein structures and fold recognition. However, many groups implemented this type of potential in protein folding simulations. One of the most familiar potentials of this type is the quasi-chemical Miyazawa-Jernigan (MJ) potential in which the residue-residue interaction matrix was derived from a statistical analysis of the number of 42 globular and 30 monomeric protein crystal structures obtained from the Brookhaven protein data bank. A statistical analysis of the probability of forming a contact with a residue was calculated and quasi-chemical approximation of the number of contacts formed gave information on the effective contact energies of the amino acid. This led to a 20-letter model representing the potential energy of the twenty amino acids respectively. This statistical analysis was actually done by fixing the proteins into a lattice model (Miyazawa and Jernigan, 1985). The work of Broglia et al. (Broglia et al., 1998) on the role of folding intermediates in the aggregation mechanisms was reviewed in section 2.3.5.3. They simulated two 36-residue peptide-model interacting via the MJ potential. This model was considered having one of the most accurate specific residue-residue interactions compared to other lattice models (Hall et al., 2006). Moreover, Leonhard and coworkers (Leonhard, Prausnitz, and Radke, 2003b) studied the
effect of interaction energies on the aggregation propensity in multiple-chain systems. Systems simulated contained from two to six lattice-chains interacting via the MJ potential. They showed that small changes in the interaction energies induce large changes in the aggregation behavior. The use of MJ potentials in protein folding and aggregation simulations are well-investigated in section 2.4.

Energy functions for the intermediate-resolution models include not only the three categories described previously but also hydrogen-bonding potentials, multibody terms, burial terms (in which the strength of the hydrophobic interaction depends on the extent of burial), and special potentials for disulfide bonds and proline. These models are used in conjunction with various highly efficient computer algorithms that constrain and guide the search through conformational space, allowing the attainment of the native state in a reasonable time frame (Hall et al., 2006).

The fourth and last category of interaction is the physics-based interaction models. Generally speaking, these types of models have connection to physics in the derivation and the functional form. Nonetheless, they are still parametrized using experimental strategies and information from experimental data banks such as the Protein Data Bank (Scheraga, Khalili, and Liwo, 2007). These physical models were pioneered by Levitt (Levitt, 1976), but the initial models were not used for proteins simulations. Many models emerged including partially physics-based ones (Wolynes, 2005; Fujitsuka et al., 2004; Liwo et al., 2001) with a knowledge-based part of the Hamiltonian, and other purely physics-models without any knowledge-based parts used. These models have been used to investigate the protein folding mechanisms and structure prediction (Fujitsuka et al., 2004). One special model is the UNRES model developed by Scheraga, Khalili, and Liwo (Scheraga, Khalili, and Liwo, 2007) which exhibits exceptional speed even when compared to all-atom simulations by folding 4000 and 200 times faster than high-resolution models MD simulations with explicit and implicit solvents respectively.
2.3.7 Solvent types

Folding and aggregation mechanisms are not only affected by residue-residue interactions, but are also governed by peptide-solvent interactions. In fact, the solvent entropy and varying interactions with the protein backbone and side-chains have significant effect on the protein dynamics. Accordingly, as with peptide-chains, the interaction of the solvent and its effect should be included in the interaction model. In this section, we present the models of solvent and solvent molecules implemented in protein simulations. As all other parameters, the level of representation of the solvent is determined by an optimization between computational cost and accuracy of representation. These models can be divided into two general categories, explicit-solvent models or implicit-solvent models.

2.3.7.1 Explicit Solvent

The most realistic representation of the solvent is by an explicit addition of the water molecules (or any other solvent molecule) inside the simulation box. This introduces a realistic environment for the peptide chains and allows the derivation of accurate kinetic and thermodynamic properties of the system (Scheraga, Khalili, and Liwo, 2007). Usually, explicit solvents are implemented in high-resolution models simulations (Hall et al., 2006), and are specially popular with experimentalists and theorists due to the availability of simulation packages with well-tested atom-atom forcefields. These include AMBER (Weiner et al., 1984; Weiner et al., 1986), CHARMM (Brooks et al., 1983), ENCAD (Levitt, 1983; Levitt et al., 1995), DISCOVER (Dauberosguthorpe et al., 1988), and ECEPP (Momany et al., 1975; Nemethy, Pottle, and Scheraga, 1983). The simulation box used with explicit-solvent models is most commonly a rectangular box but other geometric shapes of the box where also implemented (Adcock and McCammon, 2006).

The solvent can either be space filling in which case periodic boundary conditions are required, where beads hitting the wall from one side leave from the opposite wall, or another less common approach is using a single thin layer of water molecules around the peptides simulated (Adcock and McCammon, 2006). Many models for water solvents have been used in molecular dynamics simulations. These include the SPC model (Berendsen, Grigera, and
Straatsma, 1987), TIP3P, TIP4P, and TIP5P models (Jorgensen et al., 1983), the ST2 model (Stilling and RAHMAN, 1974), and the TIP4P-Ew model (Horn et al., 2004).

One shortcoming of using explicit-model solvents is the tremendous increase in computational cost due to increased number of extra degrees of freedom including water translation/rotation and internal degrees of freedom (Scheraga, Khalili, and Liwo, 2007). Moreover, water models are parametrized to a standard temperature of approximately 298 kelvin. This leads to an inaccuracy in studying the evolution of the solvent properties (like solvent densities) as a function of temperature (Horn et al., 2004). Finally, the restrictions in time scales due to such increase in computational cost allow simulations with explicit solvent to reproduce events in the order of $10^{-8}$ and $10^{-9}$ seconds for typical proteins (Day and Daggett, 2003), which is only one tenth of the the folding time of real protein (Lee et al., 2002). Accordingly, MD simulations with explicit solvents are restricted to studying a small range of events. These exclude folding pathways and the role of unfolding intermediates in the aggregation process. As a solution to this problem implicit-solvent models have become quite popular in the scientific community (Javidpour, 2012).

### 2.3.7.2 Implicit Solvent

Implicit solvent models puts forward an alternative solution to the search problem. The effect of the solvent is implicitly included in the energy function by using potentials of mean force, which are solvent-averaged interaction potentials. Many models have been developed to account for the solvent effect. Still et al. (Still et al., 1990) introduced the generalized Born surface area model (GBSA) to account for the electrostatic interactions between the solute and solvent, Vila et al. (Villa et al., 1991) expressed the free energy of solvation in terms of solvent-accessible surface areas of solute atoms, Stouten et al. (Stouten et al., 1993) and Augspurger et al. (Augspurger and Scheraga, 1996) accounted for the pair-residue excluded volume effect.

All the simulations with low-resolution models and intermediate-resolution models, like the one used by the Hall group (Hall et al., 2006) reviewed in section 2.3.5.3, use implicit solvent. Using implicit solvent has proven to be not only successful in reproducing experimental results, but effective in accessing wider time-scales.
2.3.8 Potential forms and DMD

The mathematical form of the interaction potential can be classified as either continuous or discontinuous. Continuous potentials, like the Lennard-Jones type and other soft-potential, are mainly applied in physics-based models but can also be found in the empirical-based models of proteins. A discontinuous potential provides a discretization of the interactional space which reduces the computational cost. Discontinuous potentials exert forces only when particles collide, enabling the exact - as opposed to numerical - solution of the collision dynamics (Hall et al., 2006). In this discrete potential, the forces between particles are zero or infinite. A collision occurs when two particles reach a distance at the next nearest step in their potential. The spheres are either captured or escape from the step. Since the collision of the particles drives the system to move forward in its trajectory, we refer to discontinuous molecular dynamics (DMD) as event-driven simulations. Many forms of discontinuous potential are employed including square-well potential (SW), shoulder-well, hard-sphere and many more. In this section we will focus on discontinuous molecular dynamics and how it works, then we will look at its applications and its relevance to protein simulations.

Discontinuous molecular dynamics presents a suitable form for the interaction of simplified models. This extends the accessible simulation time ($10^8 - 10^9$ times faster than regular MD) and permits rapid and accurate sampling of proteins complexes and their conformational space (Ding and Dokholyan, 2005). Larger systems can be sampled, specifically for studying aggregation in multichain systems (Ding and Dokholyan, 2005; Dokholyan et al., 2003). Unlike regular molecular dynamics which consists of solving Newton’s equations of motion, the algorithm in DMD is based on solving the ballistic equations of motion which is an iterative search of the immediate collision events in the system (Dokholyan, 2006). This reduces the computational cost even more. Moreover, MD simulations are not event-driven but rather time-driven simulations. This means that simulations are set within a time interval rather than steps or events. Continuous potentials can be mimicked by a series of square-well potentials with small steps which approximate a Lennard-Jones potential. As always there is an optimization between the accuracy of representation and the computational cost which scales with the number of steps in the potential. However, it is argued that a single step is often sufficient to reproduce generic behavior and characteristics of the
phenomena of interest (Dokholyan, 2006).

DMD was proven to be efficient in reproducing generic folding and characteristics of real proteins found in experiments. Simulations of small 20-monomer Trp-cage miniprotein showed that the native conformation folded with a root mean square deviation (RMSD) of 2 angstroms from the experimentally derived structure by NMR (Ding, Buldyrev, and Dokholyan, 2005). The validation of DMD was also true for larger systems. Ding and coworkers (Ding, Buldyrev, and Dokholyan, 2005) found the lowest energy of three 51 monomer ccb designed peptides within 1.26 angstroms of the experimentally determined native state (Kammerer et al., 2004). The fluctuations around the native state of α-lactalbumin were also studied using DMD (Vendruscolo et al., 2003). Ding and coworkers (Ding, Jha, and Dokholyan, 2005) reconciled the results of NMR spectroscopy studies on completely unfolded structures with simplified models DMD simulations, and showed the existence of significant amounts of local conformational bias toward the native state. Borreguero et al. (Borreguero et al., 2004) investigated the folding pathways of the c-Crk SH3 domain using DMD. The temperature of the simulation controls the transition states and therefore changes the folding pathways. The pathway with the smallest energetic barrier was also determined. Moreover, the work by Ding and coworkers presents another validation of DMD simulations. A system of three ccb peptides was folded with the same forcefield used for the folding of Trp-cage miniprotein (Ding, Buldyrev, and Dokholyan, 2005). The folded structures exhibited three α-helix bundles with an RMSD of 1.26 Å from the crystal structure (Kammerer et al., 2004).

Many other DMD simulations succeeded in explaining various aspects of the protein problem including fibril formation (Javidpour and Sahimi, 2011; Javidpour, Tabar, and Sahimi, 2009; Javidpour, Tabar, and Sahimi, 2008; Nguyen and Hall, 2004). Jang et al. (Jang, Hall, and Zhou, 2002a; Jang, Hall, and Zhou, 2002b) investigated the effect of temperature on equilibrium properties, folding pathways, and kinetics of single chain systems of β-strand peptides, then used these results to further investigate the aggregation phenomena of systems of four β-strand peptides. In addition, the work by Hall and Nguyen (Nguyen and Hall, 2004; Nguyen and Hall, 2006; Hall et al., 2006; Hall and Waggner, 2006) made extensive use of DMD simulations using their intermediate resolution model to investigate large systems of 48- to 96-KA14K polyalanine peptides. They showed that their model peptides,
initially set in the denatured states, formed α-helices at low concentrations and temperatures, but aggregated into fibrillar structures at high concentrations and temperatures. Also, they investigated the effect of solution conditions (including temperature and peptide concentration) and peptide-chain length on the kinetics and thermodynamics of aggregation by fibril formation.

Despite the simplicity, DMD simulations of coarse-grained models offer a unique access to the time and conformational space of complex protein simulations. Validating a model to be successful in reproducing generic behavior of protein folding requires an access to a large portion of the conformational folding space. DMD simulations using coarse-grained models is proven for such tasks and will be implemented in our study.

2.4 Development of the Off-Lattice MJ Potential Model

The off-Lattice MJ potential used in this project draws from previous studies on discontinuous potential systems and knowledge-based interaction models as well as significant advances in computational algorithms. The most relevant work to our project is the work done by the Bratko group.

Leonhard and coworkers (Leonhard, Prausnitz, and Radke, 2003a) proposed a two parameter modification of the MJ matrix to improve the protein-like behavior of model peptides through a systematic renormalization of residue-solvent interactions. This modification renders the behavior of model protein more consistent with experimental systems by optimization the resistance to aggregation, protein folding rates, and folding cooperativity.

Cellmer et al. (Cellmer et al., 2005b) investigated the competition between aggregation and refolding. They implement a coarse-grained minimalist model with each bead in the chain representing a single amino acid. The interaction potential between beads in the chain are simplified to hydrophobic, polar and neutral (e.g., implicit solvent model). Order parameters were developed to calculate the refolding yield. Also, the simulation is carried-out under experimentally relevant conditions by mapping out simulation parameters to real parameters, such as temperature and protein concentration. The pathways of refolding and aggregation were analyzed for 3 equilibrated chains starting from the denatured states and
it was concluded that using one-step dilution in large volumes for refolding has a positive effect on the refolding yield. Secondly, Cellmer and coworkers investigated the effect of ‘specific amino acids’ involved in aggregation. Although there is a lack of information related to the interactions that drive aggregation, it is thought that the exposure of a hydrophobic core induces aggregation between partially unfolded and folded chains. Some regions of the chain are more prone to aggregation by forming intermolecular interactions. This led to the idea that proteins can be rationally designed through point mutations, in order to reduce the propensity of aggregation. This is actually consistent with the protein data bank survey done by Chiti et al. (Chiti et al., 1999) who argued that evolution of proteins is associated with a reduction of long strings of hydrophobic amino acids. Therefore it can be concluded that point mutations disrupting stretches of hydrophobic amino acids might increase the refolding yield (Cellmer et al., 2005b).

The same 64-mer sequence was used in a MC simulation of lattice models (Cellmer et al., 2005a). The lattice model is validated by having an energetic bias toward the native state. In fact, the proteins sample a small fraction of the conformational space before reaching the native structure, because native interactions are generally more stable than the non-native ones. Cellmer et al. (Bratko et al., 2006; Cellmer et al., 2005c; Cellmer et al., 2005b) calculated the free energy of folding for both isolated and multichain systems. By fixing the chain sequence and length, they varied the number of neighboring molecules to monitor the competition between interprotein interaction and interaction with other molecules, and their effect on association, folding, misfolding, and aggregation. The folding/unfolding temperature shifted to lower temperatures with increasing peptide concentration. Cellmer et al. (Cellmer et al., 2005a) showed that a multichain system is more prone to misfolding and aggregation by increasing the energy barrier to reach the native state. Simulations were carried out at a reduced temperature equal to 0.1 where chains in isolation populate the native state 99% of the time. The isolated system exhibits a funnel-like energy landscape with a small energy barrier. In a two-chain system, the energy landscape becomes more rugged exhibiting two local energy minima and a small energy barrier. By increasing the number of the neighboring peptides in the system, the energy landscape becomes increasingly rugged, and the misfolded states are populated half of the time by the chains in the system and the bias toward the native state is removed. Finally, specific amino acid residue-residue contacts
were investigated. Cellmer et al. argue that for the two- and four-chain systems, 9 out of the 10 interprotein contacts found in aggregates, are buried in the native state. This indicates that unfolding induces aggregation, which is consistent with experimental results (Chiti et al., 1999).

Another study conducted by Bratko et al. (Bratko et al., 2006) looks at the effect of single point sequence alterations on the aggregation propensity of a model protein using a 64-mer on-lattice MJ model. Therapeutic proteins manipulated during biotechnological processing are more prone to aggregation due to environmental stresses. This is an incentive for protein engineers to identify the sequences that are more likely to aggregate, and eventually present an alternate sequence, while retaining the specific biological function. As a first step towards validating the preservation of biological activity due to a single-point mutation, Bratko et al. (Bratko et al., 2006) considered three 64-mer sequences where two of them have a single-point modification in their primary sequence. A comparison of native state stability is performed by looking at the interprotein contacts formed as a function of fraction of native contacts conserved. The results show that all three sequences have significant native-state population, and can therefore be regarded capable of performing the same biological function.

Secondly, as a continuation of the previous work on the competition between folding and association in multichain systems, Bratko et al. (Bratko and Blanch, 2001; Bratko et al., 2006; Cellmer et al., 2005c) argue that by increasing the number of neighbors in a system the position of the free energy minima shifts to misfolded and strongly associated states and therefore favoring interprotein interactions over the native state contacts. This effect increases with increasing peptide concentration. As a validation, two systems each containing four mutant chains with different mutations were simulated from both the unfolded and the native state. These modified sequences exhibit a very similar behavior to the original sequence as an isolated chain. Trajectories originated from the unfolded chains never refold for all three systems (original and two mutant sequences) regardless of the time of simulation. Initially each of the protein sequences fold but the modified sequences unfold and never refold. This is because the loss of native contacts are compensated by the stabilization through interprotein interactions, as the energy of the aggregate is similar to that of the native state. However, the mutated chains exhibit a strong resistance to aggregation. This suggests that
a point mutation can strongly influence the competition between folding and association. By comparing the free energy landscape of two WT chains (original chains) and two mutants, Bratko et al. (Bratko et al., 2007) observed a similarity between the effect of mutations on aggregation/folding free energy landscape to those of changing external variables such as protein concentration or temperature. This led to the conclusion that an appropriate sequence mutation can modulate the appropriate thermodynamics of aggregation and folding in a manner similar to a change in external conditions. Introducing subtle mutations allows a quantitative change of the coexistence lines of the protein solution phase diagram without changing the general behavior qualitatively.

Bratko and Blanch (Bratko and Blanch, 2001) simulated a multichain system (up to 8 chains per system) using 27-residue peptides fixed on a 3-D cubic lattice where each lattice site can accommodate a single protein. The system is initially denatured and the peptides are allowed to fold. At the end of the simulation, the chains are found to be stabilized in two types of clusters, a cluster of folded native chains, and an assembly of misfolded peptides stabilized by interprotein interactions. Bratko and Blanch observed that the association of the protein molecules in a multichain system renders the energy landscape rugged, hence slowing the configurations dynamics. This effect scales with increasing cluster size and concentration, suggesting an extrapolation to a glassy-like regime. This glassy-like regime is due to the formation of aggregates trapped in local free energy minima that exists even at high temperatures. Also, it was noted that in the folded form, peptides exhibit a strong resistance to aggregation. This shows the existence of a large energy barrier between the two favorable forms. Finally, in an attempt to mimic the function of chaperon molecules in aiding protein refolding, the effect of introducing molecular species to the systems is monitored. These molecules have a similar native structure as the peptide but with about 31% stronger intramolecular interactions. The chaperons are expected to prevent or decrease the propensity of interprotein interactions by shielding proteins from each other and therefore promoting folding. The improvement in the refolding yield was noted by an increase in the folded population densities in 11 out of 15 runs (Bratko and Blanch, 2001).

Moreover, Cellmer et al. (Cellmer et al., 2005c) simulated two 64 monomer on-lattice chains
to study the effect of interprotein interactions on conformational change. The melting temperature occurs at a reduced temperature of 1.15, which is consistent with that reported previously by Leonhard et al. (Leonhard, Prausnitz, and Radke, 2003a) and shows that implicit representation of the solvent in the potential function is equivalent to the explicit representation used by Leonhard et al. (Leonhard, Prausnitz, and Radke, 2003a). At the lowest temperature, chains are mainly dominated by native contacts. The unfolding transition shifts to lower temperatures with addition of a peptide. This conclusion was verified by looking at the radius of gyration which shows that for a two-chain system the peptides expand at lower temperatures more than chains simulated in isolation. This unfolding is accompanied by a ‘trading’ of intramolecular energy with intermolecular energy.

Bratko et al. (Bratko et al., 2007) show that coarse-grained models of proteins play a major role in the conceptual understanding of protein folding. The specific heat capacity computed in the simulation with an appropriately chosen energy scale, is consistent with the heat capacity measured experimentally by calorimetry. Moreover, the data collected from computer simulation are consistent with those observed in protein denaturation experiments using intrinsic tryptophan fluorescence or circular dichroism (Bratko et al., 2007). The melting temperature obtained is 310 K, and the energy of unfolding is found to be 28 Kcal/mol which are both consistent with experimental data. They also suggested that there are ‘hot sites’ which can be mutated in order to decrease the propensity of aggregation. The simulation of a multichain 64-mer on-lattice system indicates that a stretch of 12 amino acids controls the aggregation process. Several of these amino acids are characterized with high hydrophobicity and are therefore buried in the core of the native state. With the partial unfolding of the peptide, these amino acids become exposed to interprotein interactions thereby increasing the propensity of aggregation, which is consistent with reports of some experimental results (Fink, 1998). Other amino acids, specifically ‘charged residues’ found on the surface of the folded peptide, affect aggregation by interacting with complementary ‘charged residues’ found in neighboring proteins.

The protein refolding problem is investigated (Bratko et al., 2007) by looking at an off-lattice model that folds into a β-barrel structure previously studied by Honeycutt and Thirumalai (Honeycutt and Thirumalai, 1992). In this model, the peptide consists of three types of amino-acids: ‘hydrophilic’ beads with a repulsive potential, ‘hydrophobic’ beads interacting
via a Lennard-Jones potential, and ‘neutral’ beads or ‘glycine’ beads. Three chains were equilibrated without attractive interactions, when the attraction is turned-on the system was monitored until all the chains were either associated or folded. The number of refolding yields was calculated and it was observed that this yield increases three times (from 15 to 51%) when the effective peptide concentration decreases from ∼ 3 mM to ∼ 0.3 mM (volume increases) (Bratko et al., 2007). Similarly, the specific contribution of different strands to the aggregation propensity was investigated. Two specific strands, each with a replica, were simulated in a 4-chain system. Although the strands have the same number of hydrophobic residues, they exhibited different propensity for aggregation. A direct implication is that the amino-acid sequence cannot be used alone to identify the aggregation-prone segments. As a further testament to this argument, several mutant chains are designed, where two hydrophobic residues are replaced by hydrophilic residues without altering the native state or stability of the peptide model. Mutation to the strand which is more prone to aggregation increases the refolding yield at a higher degree. This is in good agreement with previous analysis. The fact that aggregation is controlled by a small number of specific amino acid is consistent with experimental work, where amyloid fibril are formed by unfolded proteins (Chiti et al., 2002).
Chapter 3

Research Objectives

The very broad title of this study, ‘Understanding biopharmaceutical aggregation using minimalist models based on square-well potential’, comprises results from simulating two loosely-related systems with discontinuous dynamics: one simulation involves chains composed of spheres interacting through square wells, while the other one is mono-dispersed spheres interacting through square-well potentials. The main objective of the first project is to understand the causes and mechanisms of protein aggregation, with the ultimate aim of preventing it. In the second project, the goal is to characterize the supercritical behavior of a SW fluids, with special focus on the adhesive model which represents protein solution with short-ranged potential.

Most of the protein systems simulated using the Miyazawa-Jernigan potential studied aggregation from the non-native or denatured states with lattice models. In reality, neither the amino acids of proteins are fixed on a lattice, nor aggregation-phenomena of interest occur from the unfolded-states with a biasing towards native structures. Therefore we put forward a study on the aggregation from the native state by simulating a multichain system of proteins as an off-lattice residue interacting via the MJ model. In the second project we investigate a new empirical method for calculating the critical temperature of SW fluids, specifically short-ranged systems which are especially difficult to simulate as they exhibit glassy behavior. Other methods require using simulations at lower temperature in the two phase region which are more difficult to equilibrate.
3.1 Effect of model parameter: finding a good model for studying protein behavior

The first part of this project consists of showing that our model can be used to study folding and aggregation phenomena. Like any other model, many parameters can be fine-tuned to change the characteristics of the system displayed, which is the behavior of the peptide while folding in this case. We aim at choosing the best model (with the selected set of parameters) which exhibits the generic behavior of ‘real’ proteins. This is quite a broad statement but the main features of protein folding are well known. These include:

- An unfolded structure at high temperature
- A well-defined unique structure at low temperature
- A single-step or a two-step structural transition from the unfolded to the folded state with decreasing temperature (consisting of the collapse and freezing transition)
- A large enough core of amino acids that are inaccessible to the solvent

These characteristics are a summary of the thermodynamic aspect of the protein folding problem. Effectively what we are looking for is a model-chain that exhibits a single step or a two-step transition with decreasing temperature that forms a unique, well-defined ensemble of structures. In protein computational models, the parameters that can be manipulated are the coarse-graining level, the form of the interaction potential, the sequence of residues, the length of the chain equivalent to number of beads ‘N’, the overlap parameter, defined as the ratio of the hard-core diameter to bond length, which refers to the chain’s stiffness ‘σ/l’, and the range of interaction potential ‘λ’. The simulator DynamO is an event-driven package which allows the simulation of simplified models interacting through discontinuous potentials. Accordingly, the model chosen is a simplified coarse grained model where each bead represents a single amino acid and the MJ model is used to represent the interaction energies between all possible pairs of amino acid types. Although DynamO can be used for simulating models with simplified square-well potentials and semi-continuous potentials (e.g., stepped potentials), we chose to use a single SW potential to reduce the computational cost. This allows tracking the exact formation/breaking of the contact-bonds between residues and allows us to define a contact map, which provides a representation
for all the non-covalent bonds stabilizing the protein structure. As shown previously, the work pioneered by the Bratko group was the starting point of this study. They simulated multichain systems using lattice models interacting via the MJ potential. Many parameters in our model will be set accordingly in the initial models simulated. Our simplified model uses the MJ potential and the same sequence adopted by Bratko and coworkers, however chains can move freely in the box, as they are not fixed on a lattice.

Accordingly the first objective is to investigate the effect of the chain length \( N \), the range of interaction potential \( \lambda \), and the overlap parameter \( \sigma/l \) on the folding characteristics, then fine-tune these parameters to show that our model can exhibit the generic behavior of real proteins.

Bratko et al. (Bratko et al., 2006) argued that for the MJ on-lattice model, the 64-mer chain is the smallest chain able to reproduce satisfactorily the folding behavior of ‘real’ proteins. Accordingly, the first step was to look at the folding behavior of a 32-mer and a 64-mer chains. The cooperativity and the collapse of a big core are known to be characteristic of proteins behavior which are well established for long chains. Smaller chain on the other hand are easier to equilibrate and were seen to exhibit two-stage folding. So the choice of the chain length is a compromise between structural characteristics and computational cost. We investigate the shortest chain capable of reproducing folding characteristics of real chains.

We aim at evaluating the effect of the overlap parameter \( \sigma/l \) on the structures and transitions as a function of temperature. Many questions are investigated: do chains with different stiffness fold according to different pathways? Do they exhibit similar structural transition? Do they have different intermediate stages? Which folding behavior is closer to that of a protein?

It is argued that changing the interaction width \( \lambda \) has a tremendous effect on the folding pathways (Taylor, Paul, and Binder, 2009a; Ruzicka, Quigley, and Allen, 2012). By decreasing the range of interaction potential, homopolymers shift from a two-step folding pathway, to a discontinuous single-step folding, in which the ‘extended coil’ collapses directly into the native state with decreasing temperature. Accordingly, our next objective is to investigate the effect of \( \lambda \) on the folding behavior of our heteropolymer model, and find the best range
of interaction value which exhibits a one-step or two-step folding.

Once the effect of parameters is studied, we will choose the ones corresponding to models that best represent the generic aspects of protein structure.

In summary the first part of the project consists of investigating the effect of model parameters on the different structural transitions. Afterwards, theses parameters are fine-tuned to get the best set for an optimized model which exhibits generic behavior of ‘real’ proteins. This model will be used for further analysis.

3.2 Low-temperature structure: a new order-parameter to define the native ensemble

At low temperatures chains exhibit small conformational fluctuations or ‘breathing motions’. This flexibility is crucial for some protein’s catalytic activity. On the other hand, at high temperature chains exhibit large conformational fluctuations due to structural transitions. These fluctuations can be correlated with unfolding and/or aggregation. These observations provide the next objective, which is to identify the nature of these two types of fluctuations, and then investigate whether or not the conformational breathing at low temperature correlates with the structural transitions at high temperature.

The first step is to characterize the low-temperature and the high-temperature behavior, which allow us to define the native conformation of the chain. Many definitions exist in literature (Tang and Dill, 1998; Cellmer et al., 2005c; Cellmer et al., 2005b) but mainly the native state is either considered as a single stagnant conformation, or an ensemble of fluctuating conformations. Once the native state is characterized, the next task of the project consists on finding a distinctive property which discriminates the conformational space between folded and unfolded structure by separating low-temperature breathing motion and high temperature unfolding and aggregation. This definition should agree with either the freezing or melting/collapse temperature. In this section we aim at answering the following questions: what is the nature of fluctuations at low temperature? How do we differentiate between partial unfolding, which promotes aggregation, and conformational fluctuations, which is important for protein’s catalytic activity?
3.3  Investigating the correlation between the behavior of isolated peptides and aggregation propensity in multichain systems

There is a strong experimental correlation between the protein melting temperature and aggregation at room temperature. The implication is that aggregation occurs through the ensemble of unfolded states, a lower melting temperature implies a higher fraction of unfolded states at room temperature, and hence a higher likelihood of aggregation. However, at room temperature the fraction of proteins in the unfolded state will be small. Another possibility is that the melting temperature correlates with some other property at room temperature, which determines the aggregation propensity. The core of the chain appears as a good candidate due to its role in the behavior of the chains. Going from high to low temperatures, unfolded chains collapse into compact structures due to the formation of a hydrophobic core. At slightly elevated temperature, or upon a change in solution conditions, many chains in their native state partially unfold and aggregate by exposure of their attractive cores. Thus there might be a strong correlation between aggregation propensity and the folding temperature, implying that the behavior of the protein at high temperatures correlates with the properties at room temperature. Understanding this behavior requires mapping the reduced temperatures in modeling experiments to physical temperature in the real world. So what is the role of the core in folding behavior and aggregation propensity of protein chains? Can we use the properties of the core to define a native-ensemble of states, and hence divide all structures in the folding pathway between folded and unfolded allowing us to calculate energy difference between states? How does the flexibility of the chain at low temperature correlate with its stability at the folding temperature and the aggregation propensity?

3.3  Investigating the correlation between the behavior of isolated peptides and aggregation propensity in multichain systems

The objective here is to investigate the existence of a correlation between the behavior of chains in isolation and in multichain systems by examining the aggregation behavior in terms of the transition temperatures. More precisely, we seek to determine what properties of the isolated chains correlate with the aggregation propensity in multichain systems? What intermediates are aggregating and are there conformational changes upon aggregation? We present results of simulations containing up to 8 chains for 5 different peptide
models with varying sequences.

In conclusion, the aims of the first project can be summarized as follows:

- Characterizing the behavior of the protein model for a 32mer and a 64mer, in order to identify a short enough chain capable of reproducing the generic behavior of ‘real’ proteins.

- Evaluating the effect of stiffness, the range of interaction potential, and chain length on the protein structural transitions.

- Choosing the best model to study protein folding behavior and aggregation.

- Define an order parameter or property which divides the conformational space between folded and unfolded structures.

- Study the correlation between low-temperature fluctuations and high-temperature conformational transitions.

- Evaluate the relationship between structural properties of isolated chains and their aggregation propensity in multichain systems.

### 3.4 Critical parameters calculation in square-well fluids

In the second project, we draw an analogy between the effect of the range of interaction on the folding behavior of an isolated chain, and the phase behavior of a SW fluid. This analogy was previously investigated for homopolymer with very simplified models (Dobson, Sali, and Karplus, 1998). Given the relevance of the protein phase behavior to many bioprocessing phenomena, such as crystallization and aggregation, we simulate isotherms in the supercritical regions of the phase diagram of a SW fluid of \( N = 10976 \) particles for \( \lambda \) equal to 1.005, 1.2, and 1.5. Therefore, we describe the behavior of square well fluids with different range of interaction potentials with special focus on strongly short-ranged systems with \( \lambda \) equal to 1.005, collectively known as the sticky sphere limit. Here we introduce a more efficient way of calculating the critical temperature based on experimental observation specifically for strongly short-ranged systems which are especially difficult to simulate, as they exhibit glassy behavior. Other methods require using simulations at lower
temperature in the two phase region and hence will be more difficult to equilibrate. This method could then be applied to more realistic short-ranged potentials for proteins which could be patchy spheres or possibly non-spherical shapes interacting through anisotropic interactions. Identifying the critical temperature and how it depends on the pair potential is very important in formulations as we need to predict when the solution will go opalescent.
Chapter 4

Approach

4.1 Model Description and Methodology

4.1.1 Model

In this project, all the simulations are performed using a modern event-driven molecular dynamics simulations package, DynamO, developed by Marcus Bannerman, and distributed under the GNU General Public License. The full source code and documentation are freely available online at http://dynamomd.org/. DynamO provides reference implementations of the best available event-driven algorithms. These techniques allow the rapid simulation of both complex and large systems.

DynamO performs molecular dynamics simulations in the canonical ensemble (NVT). The microcanonical ensemble (NVE), which fixes the number of atoms, the volume of the periodic box and the total energy of the system, is coupled with the Andersen thermostat for simulating a canonical ensemble at a constant temperature (Liu, Garde, and Kumar, 2005).

In this study, the protein is modeled as a square well, off-lattice heteropolymer. The chain consists of \( N \) linearly bounded hard spheres, where each sphere represents a single amino acid. The model is self-avoiding, meaning if a bead is trying to move into a place occupied by another bead, the move is automatically rejected. The beads interact via quasi-chemical Miyazawa-Jernigan square-well potential, following the pioneering work computed by Bratko et al. (Bratko et al., 2007).
4.1.2 Simulation Details

Simulations are performed on the university of Manchester computer clusters csf.itservices.manchester.ac.uk and tom.ceas.manchester.ac.uk, which are 8 core, 2 GHz Intel computers. This allows the use of parallel tempering and multiple copies to overcome the energy barriers, allowing the simulation of larger systems with reasonable time scales.

In molecular dynamics simulations, the computational intensity is a function of three major tasks: event detection (i), maintenance of the event list (ii), and execution of the event (iii). Usually, these tasks lead to an $O(N \log N)$ scaling with the number of particles $N$ of the computational cost in most standard algorithms, DynamO on the other hand includes several new algorithms for mitigating round-off error, improving speed, and optimizing memory access patterns, allowing it to display the optimal $O(N)$ asymptotic scaling of the computational cost.

4.1.3 Interaction Potentials

The protein is modeled as freely jointed chain of $N$ hard-sphere monomers with diameter $\sigma$, bond length $l$, and flexibility $\delta$. The latter parameter controls the allowed fluctuations in the bond length of plus or minus $\delta$. A value equal to 0.1 is used here based on simulations studies of AB heteropolymers by Magee and coworkers (Magee, Warwicker, and Lue, 2004) allowing 10% fluctuation of the bond length. Magee and coworkers argued that this parameter has no effect on the heteropolymer folding results. However, having some bond fluctuations allows for faster equilibration or improved sampling efficiency.

Bonded monomers $i$ and $i + 1$, separated by a distance $r$ such as those represented in Figure 4.1 (a), interact via a potential defined as follows:

\[
\begin{align*}
  u_{i,i+1}(r) &= \begin{cases} 
    \infty & \text{for } r < (l - \delta), \\
    0 & \text{for } (l - \delta) < r < (l + \delta), \\
    \infty & \text{for } (l + \delta) < r.
  \end{cases}
\end{align*}
\]
If \( i \) and \( j \) are not directly-bonded monomers, they will interact via square-well potential, dependent on the distance \( r \) as shown in fig 4.1(b):

\[
\begin{align*}
  u_{i,j}(r) = \\
  \begin{cases}
    \infty & \text{for } r < \sigma \\
    -\epsilon & \text{for } \sigma < r < \lambda\sigma \\
    0 & \text{for } \lambda\sigma < r
  \end{cases}
\end{align*}
\]  

(4.2)

Reduced scale units are utilized when simulating in DynamO. These are functions of hard-sphere diameter \( \sigma \), temperature of the system \( T \), Boltzmann’s constant \( k_B \), and characteristic
energy \( f(\sigma, T, k_B, \epsilon) \):

\[
\text{Dimensionless properties:} \quad \begin{align*}
    r^* &= \frac{r}{\sigma} \\
    E^* &= \frac{E}{\epsilon} \\
    T^* &= \frac{k_B T}{\epsilon} \\
    C_v^* &= \frac{\langle E^2 \rangle - \langle E \rangle^2}{T^2} 
\end{align*}
\]

(4.3)

All properties reported in the dissertation are in dimensionless form, hence the \(^\ast\) is omitted.

The MJ model considers the relative hydrophobicity of all 20 amino acids. As such, a 20x20 matrix is required to describe the interaction between any pair of non linearly-bonded beads, Table 4.1 shows the MJ potentials used to construct this matrix.

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Symbol</th>
<th>Potential</th>
<th>Type</th>
<th>Amino Acid</th>
<th>Symbol</th>
<th>Potential</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenylalanine</td>
<td>F</td>
<td>-6.85</td>
<td>H</td>
<td>Glycine</td>
<td>G</td>
<td>-2.17</td>
<td>H</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>I</td>
<td>-6.33</td>
<td>H</td>
<td>Threonine</td>
<td>T</td>
<td>-1.72</td>
<td>P</td>
</tr>
<tr>
<td>Methionine</td>
<td>M</td>
<td>-6.06</td>
<td>H</td>
<td>Asparagine</td>
<td>N</td>
<td>-1.59</td>
<td>P</td>
</tr>
<tr>
<td>Leucine</td>
<td>L</td>
<td>-5.79</td>
<td>H</td>
<td>Serine</td>
<td>S</td>
<td>-1.48</td>
<td>P</td>
</tr>
<tr>
<td>Cysteine</td>
<td>C</td>
<td>-5.44</td>
<td>P</td>
<td>Arginine</td>
<td>R</td>
<td>-1.39</td>
<td>P</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Y</td>
<td>-5.42</td>
<td>H</td>
<td>Glutamine</td>
<td>Q</td>
<td>-1.18</td>
<td>P</td>
</tr>
<tr>
<td>Valine</td>
<td>V</td>
<td>-4.94</td>
<td>H</td>
<td>Proline</td>
<td>P</td>
<td>-1.18</td>
<td>H</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Y</td>
<td>-3.55</td>
<td>P</td>
<td>Aspartic acid</td>
<td>D</td>
<td>-0.96</td>
<td>P</td>
</tr>
<tr>
<td>Histidine</td>
<td>H</td>
<td>-2.78</td>
<td>P</td>
<td>Glutamic acid</td>
<td>E</td>
<td>-0.89</td>
<td>P</td>
</tr>
<tr>
<td>Alanine</td>
<td>A</td>
<td>-2.51</td>
<td>H</td>
<td>Lysine</td>
<td>K</td>
<td>0.13</td>
<td>P</td>
</tr>
</tbody>
</table>

**Table 4.1**: MJ potentials of all 20 amino acids is ascending order. The type of amino acids is referred to with the letters 'H' for hydrophobic amino acids and 'P' for polar ones.

### 4.2 Methodology

In the first project, the simulations are performed using parallel tempering sampling method.

#### 4.2.1 Temperature calculation

The set of temperatures used for parallel tempering must be defined prior to simulation. Various methods in literature are available to define the initial set of temperatures. In our work, if a run is performed for the first time, temperatures are chosen using a power law, in which the ratio between consecutive temperatures is a constant, \( A \), that is:
4.2. Methodology

\[ \frac{T_{i+1}}{T_i} = A \]  

(4.4)

The highest temperature \( T_{\text{high}} \) needs to be chosen to allow for rapid equilibration and minimize energy barriers. Once the number of temperatures \( N \) is chosen, the constant \( A \) can be calculated from

\[ \frac{T_{\text{high}}}{T_{\text{low}}} = A^{N-1} \]  

(4.5)

The initial values of \( T_{\text{low}} \) and \( T_{\text{high}} \) are set based on the work by Bratko et al. (Bratko et al., 2007; Bratko et al., 2006). Then these temperatures are iterated based on the simulation data analysis.

4.2.2 Temperature optimization

In parallel tempering simulations, temperatures of different replicas need to be optimized in order to insure that the configurations are swapped efficiently between temperatures to reach the lowest energy state. Details of various temperature optimization models and acceptable swapping rate are available in literature, and briefly discussed in section 2.3.4. Based on the findings of Kofke (Kofke, 2002), temperatures are chosen such that the acceptance rate, which is the overlap between neighboring energy histograms, is between 10\% and 30\%. This is done using energy histograms and the data on the replica exchange, including the number of attempted swaps at each temperature and the fraction of accepted swaps. Once the set of replicas chosen at various temperatures satisfies the overlap set, increasing the number of replicas would only increase the computational intensity of the simulation.

4.2.3 Data Analysis

The main data analysis tools used here include energy histograms, the contact maps, the radius of gyration \( (R_g) \), reweighted plots of energy versus temperature (rewighted with Weighted Histogram Analysis Method - WHAM) \( (E^*) \), the dimensionless heat capacity \( (C_v^*) \).
Other tools we developed including the surface accessible (SASA), the root mean square deviation (RMSD), the root mean square deviation in contact maps of the chain (RMSD<sub>CM</sub>) and in contact maps of the core residues (RMSC). A detailed description of the data and their implementation follows.

- **Contact Maps**: This is an N x N matrix, where N is the number of beads in the chain. This matrix, plotted in 2D, shows the contacts between any pair of monomers when the corresponding pairs of spheres are located within the square well separation. The monomers are presented on the two axis, in correlation with their position in the chain. In a contact map corresponding to a single snapshot or single structure, a black square denotes that no contact is made, a white square shows that the pairs are in contact. In contact maps representing an average over many snapshots or configurations, a gray square indicates that the pair of monomers are in contact for part of the simulation, with the intensity of the color inversely related to the fractional occurrence the contact is made over the configurations being sampled.

- **Radius of gyration, \( R_g^* \)**: This property is defined as the root mean square of the distance of all atoms from the center, representing a measure of the size and the 3-dimensional shape of the heteropolymer. It is calculated from the individual three spatial dimensional radii \( R_1, R_2, \) and \( R_3 \) such as:

\[
R_g^* = (R_1^2 + R_2^2 + R_3^2)^{1/2}
\]  

The principal axes \( R_1, R_2, \) and \( R_3 \) are the main axes of a member which are perpendicular and intersect each other at the center of area or ‘centroid’. One can think of the centroid as the center of gravity of an object if it is made of one material (homogeneous). Here we define 1 as the largest principal axis, 2 as intermediate, and 3 as the smallest principal axis.

- **Dimensionless internal energy, \( E^* \)**: The internal energy is the sum of square-well interactions between pair of monomers. It is defined in equation 4.3. DynamO uses the weighted histogram analysis method to generate a plot of energy versus temperature using the data generated from all the replicas simultaneously.
4.2. Methodology

- **Dimensionless specific heat capacity, \( C^*_v \):** This data type gives insight on the structural transitions in the folding pathway of the heteropolymer. The heat capacity is determined from the reweighting procedure. A structural transition into a more ordered state is correlated with a peak in the heat capacity. This allows identifying, in complementary with other tools, the major structural transitions in the folding process.

- **Root mean square deviation, RMSD:** The RMSD reflects the similarity in space between a pair of chains. The RMSD between chain \( a \) and chain \( b \) is defined as

\[
\text{RMSD}_{a,b} = \sum_i (r_{i,a} - r_{i,b})^2
\]

(4.7)

where the subscript \( i \) corresponds to the sphere label (or position of the sphere in the sequence). Thus the subscript \( i, a \) corresponds to the \( i \)th sphere in structure \( a \). The left hand side of equation 4.7 is just the squared distance between the same sphere \( i \) in two different structures \( a \) and \( b \):

\[
(r_{i,a} - r_{i,b})^2 = (x_{i,a} - x_{i,b})^2 + (y_{i,a} - y_{i,b})^2 + (z_{i,a} - z_{i,b})^2
\]

(4.8)

where \( x, y, \) and \( z \) have their usual meaning as the labels of the coordinate system. The value of \( \text{RMSD}_{a,b} \) is calculated after a mathematical routine is used to minimize the difference in the RMSD of structures \( a \) and \( b \).

In this work, we developed an algorithm to cluster structures according to their RMSD values. In this algorithm, the RMSD is calculated between all possible pairs of structures. Then for each target structure, the number of neighbors are calculated where a neighbor is defined as a structure which has an RMSD less than a specified cut-off value. The target structure which has the most neighbors becomes a representative structure and the set of all neighbors forms a cluster. The representative structure and the neighbor structures are then removed from the structural set, and the process is repeated to determine the next largest cluster. This process is repeated until no more clusters exist with sizes greater than a specified minimum value.
• **Root mean square deviation in contacts, RMSD\textsubscript{CM}:** The RMSD\textsubscript{CM\textsubscript{a,b}} reflects the root mean square deviation between a pair of contact maps \textit{a} and \textit{b}. It is defined as

\[
RMSD_{CM}^{1/2} = \frac{\sum_{i,j} a_{i,j} b_{i,j}}{\sqrt{n_a n_b}}
\]  

(4.9)

The sum is over all the elements to the left of the diagonal on the contact map matrix. Here \textit{a}_{ij} and \textit{b}_{ij} correspond to the \textit{i, j} element of the contact map matrix for structure \textit{a} and \textit{b} respectively. These elements are either 0 if no contact is formed between sphere \textit{i} and sphere \textit{j} or 1 if a contact is made between sphere \textit{i} and sphere \textit{j}. If an \textit{i, j} contact is conserved between structures \textit{a} and \textit{b}, then \textit{a}_{i,j} \textit{b}_{i,j} is equal to 1, in all other cases, the product is equal to 0. When structure \textit{a} is the same as structure \textit{b} and all the contacts are conserved, then RMSD\textsubscript{CM\textsubscript{a,b}} is equal to 1. Thus the RMSD\textsubscript{CM\textsubscript{a,b}} corresponds to the fraction of contacts that are conserved between the two structures, as structures become more different the value of RMSD\textsubscript{CM\textsubscript{a,b}} decreases.

• **Solvent accessible Surface area, SASA:** SASA is a measure of the surface area of the heteropolymer, which is accessible to the center of a theoretical solvent molecule with a specified radius. A residue with low SASA indicates that it is buried inside of the chain, and cannot be accessed by a solvent, however a high SASA indicates that the relative residue is on the surface of the chain.

Deciding the cutoff in SASA value below which the residues are considered to be buried is somehow subjective and depends on many variables. In theory, a cutoff equal to zero appears as the best choice. From experiments, for such a cutoff many residues will have very low SASA values (equal to 1,2, ..., etc) but it is not very obvious that such residues can be considered to be accessible by the solvent. Therefore, the cutoff is set in a way to have a discontinuity in SASA values between all residues, i.e., residues will have a SASA equal or slightly smaller than the cutoff, or a SASA value that is much higher. Secondly, we try to choose the cutoff such that the number of ‘buried’ residues at different temperatures is consistent with that found in literature. Choice of SASA for every chain is outlined in the model sections respectively.

• **Root mean square deviation in contacts of the core, RMSC:** The RMSC reflects the root mean square deviation between core residues defined by the SASA analysis. It is
calculated is the same way as the $\text{RMSD}_{CM_{a,b}}$. 
Chapter 5

Results and Discussion: effect of model parameters

In this chapter, we search for a set of parameters, mainly the number of residues ‘N’, the overlap parameter defined as the ratio of hard core diameter to bond length ‘σ/l’, and the range of interaction potential ‘λ’, which exhibit a generic behavior of protein chains. The characteristics we aim at finding include a high-temperature collapse transition forming a compact structure with a buried core, and a low temperature well-defined native-state with certain resistance to temperature change and a well-buried stable core. Accordingly, we investigate the effect of each parameter on the behavior of our model.

First, we explore the effect of size of the chain for various models in order to find a short-enough chain capable of reproducing generic behavior of proteins. We present results of a 32-mer chain, a 64-mer chain, and longer chains (128-mer and 256-mer). In these simulations we introduce the main analysis tools employed to identify every transition and intermediate exhibited during the folding as a function of temperature. Many 32-mer model chains were simulated with varying the range of interaction potential and overlap parameters while keeping the chain sequence and number of residues constant. However, all systems were lacking distinctive characteristics such as a stable core at low temperature, and were hard to equilibrate. Results of these simulations are not reported here, because the exact behavior of the chain is not pertinent to our discussion. Instead we show results of one selected chain to show the folding behavior of short models representing small proteins.
First, we present detailed results for a 64-mer exhibiting major characteristics found in protein behavior before discussing the effects of model parameters. Next, we investigate the effect of the overlap parameter or stiffness on the folding transitions and the low-temperature behavior. We initially present results of a system of different peptides where we fix the chain length and the range of interaction and change the overlap parameter in each peptide. Next, we repeat the same analysis for a set of peptides with a different value of $\lambda$ in order to show that the effect of $\sigma/l$ extracted is independent of all parameters. In the last chapter, and after choosing the optimized chain length and overlap parameter, we examine a series of 7 simulations with varying range of interaction potential. Models are identified that capture generic behavior of protein chains and are optimized to minimize the computational intensity.
5.1 Effect of number of beads ‘N’

This section is divided into 3 main parts: the folding of a 32 monomer chain, the folding of a 64 monomer chain, and the folding of longer chains. In each study we first give a short overview of the model and the methodology used, followed by an overview of the main results and a short discussion. From the simulation results we identify the main transitions, the intermediates, and the nature of the low-temperature structure. We finally conclude with the general folding behavior (single-step folding, two-step folding, multiple transition...), and compare it with those reported in literature. Using the data from the three models with varying length, we identify the optimum number of residues to use in later studies.

5.1.1 Folding behavior of a 32 monomer chain

The model-chain simulated is a 32 monomer chain with a short range of interaction potential $\lambda = 1.03$ and overlap parameter $\sigma/l = 1.3$. In this section we investigate the transitions and intermediates in the folding pathway, and identify the nature of the low-temperature structure with the ultimate aim of showing whether or not the 32-mer model exhibits generic features of real proteins folding.

5.1.1.1 Model

It was previously suggested that coarse-grained model proteins interacting via the Miyazawa-Jernigan potential with less than 64 beads do not exhibit generic behavior of real proteins (Bratko et al., 2007; Leonhard, Prausnitz, and Radke, 2003b). To that end, various 8-mer, 16-mer and 32-mer models were simulated with a range of interaction potential, stiffness, and sequence. The sequences are chosen by satisfying the probability of finding every type of amino acids in real proteins. Indeed, all models simulated did not exhibit generic behavior and characteristics of protein folding. Here we present simulation of one off-lattice 32-mer model chain with residues interacting via the MJ potential in order to show the characteristics of folding of small proteins. Simulation have been performed using replica exchange. Several runs using the same sequence with different starting configurations are performed in order to determine whether or not equilibration is reached. A detailed description of the
model, interaction potential, and analysis tools common to all the simulations have been introduced in chapter 4. Here we describe the key input parameters and specific simulation details related to this section.

- **Sequence**: PTSTNHMQVLCIFYPALDAEGKESGRVKW

- **Other parameters**: We use a range of interaction potential $\lambda$ equal to 1.03, which is the shortest range of interaction used in this project, and an overlap parameter $\sigma/l$ equal to 1.3.

- **Equilibration run**: The system was equilibrated after 20 million time-steps to the lowest temperature.

- **Production run**: Two independent runs of 10 million time-steps each were used for data analysis.

- **Temperature list**: In table 5.2, we present the set of temperatures of each replica in the simulation. Temperatures are in reduced units.

<table>
<thead>
<tr>
<th>List number</th>
<th>Temperature</th>
<th>List number</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.040</td>
<td>9</td>
<td>0.212</td>
</tr>
<tr>
<td>1</td>
<td>0.046</td>
<td>10</td>
<td>0.261</td>
</tr>
<tr>
<td>2</td>
<td>0.053</td>
<td>11</td>
<td>0.31</td>
</tr>
<tr>
<td>3</td>
<td>0.071</td>
<td>12</td>
<td>0.383</td>
</tr>
<tr>
<td>4</td>
<td>0.083</td>
<td>13</td>
<td>0.455</td>
</tr>
<tr>
<td>5</td>
<td>0.105</td>
<td>14</td>
<td>0.556</td>
</tr>
<tr>
<td>6</td>
<td>0.128</td>
<td>15</td>
<td>0.666</td>
</tr>
<tr>
<td>7</td>
<td>0.144</td>
<td>16</td>
<td>2.092</td>
</tr>
<tr>
<td>8</td>
<td>0.178</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**5.1.1.2 A general overview**

The energy histograms of 17 reduced temperatures is displayed in figure 5.1. The overlapping of the energy histograms from the lowest to the highest temperature indicates that structures from various temperatures swap efficiently. A detailed description on the replica exchange is available in sections 4.2.1 and 4.2.2. In order to have an accurate quantitative analysis of the data collected during the simulation, the system should be equilibrated. We
consider the system to be equilibrated if the specific heat capacity plots from two runs of the same system overlap with each other.

![Energy histograms](image)

**Figure 5.1:** A plot of the energy histograms of the 32-mer with $\sigma/l$ equal to 1.6 and $\lambda$ equal to 1.03.

This is clear in figure 5.2 which shows plots of the specific heat capacity of the system as a function of reduced temperatures taken from two independent production runs of 10 million time steps. We choose to use specific heat capacity as an indicator of equilibration as fluctuation properties are much more sensitive than averaged properties to equilibration. The heat capacity curve in figure 5.2 contains two well-defined major peaks: the first peak at $T^* = 0.18$ and a second peak at $T^* = 0.07$. The peaks in heat capacity plots are identified as folding transitions with changing temperature (Kolinski, Godzik, and Skolnick, 1993; Kaya and Chan, 2000). Accordingly, the folding of the 32-mer exhibits two defined transitions, the high-T transition starts at temperature 0.26 and exhibits a predominant peak at 0.18 and a dominant peak at 0.15 while the second peak starts at 0.105 and peaks at 0.07.

Figure 5.3 shows the variation of reweighted internal energy plotted with reduced temperature for the corresponding runs in figure 5.2. At a high temperature, approximately between $0.26 < T^* < 2.2$, the reweighted internal energy is almost constant around zero and non-bonded spheres do not interact with each other. At temperature around $T^* = 0.26$, the
Chapter 5. Results and Discussion: effect of model parameters

Figure 5.4 is a plot of the radius of gyration in terms of the three individual principal components $R_1$, $R_2$, and $R_3$. The $R_g$ plot provides insight on the 3-dimensional shape and size of the peptide. At high temperature, $R_g$ is constant at the largest value. Going to lower temperatures, at $T^* = 0.26$, there is a sharp decrease in the $R_g$ value, shown by the steep slope of the curve. At low temperature, the $R_g$ reaches a very low value, which indicates that the chain adopts a compact conformation.

The large value of $R_g$, coupled with the zero value of internal energy at high temperature range confirms that the heteropolymer is in an extended state. Going to lower temperatures, at $T^* = 0.26$, the sharp decrease in the $R_g$ value occurs at the same temperature where the internal energy starts decreasing and at the temperature corresponding to the
5.1. Effect of number of beads ‘N’

high-temperature peak in the heat capacity plot as shown in figures 5.4, 5.3 and 5.2. The initial overview shows that at high temperature no contacts are formed between residues and the peptide model is an extended state which collapses into a compact structure at low temperature with low energy due to multiple monomer-monomer interactions. This behavior occurs through a series of two transitions. The nature of these transition cannot be identified at this point but one possibility for the two peaks in the heat capacity plot is a collapse transition followed by an ordering transition. Many studies related to homopolymers (Ruzicka, Quigley, and Allen, 2012; Taylor, Paul, and Binder, 2009a; Taylor, Paul, and Binder, 2009b) argue that transitions upon which peptides collapse into ordered structures are identified as the high-temperature collapse transition where the random coil form a compact ‘globule’, and a low-temperature freezing transition towards the native state where the chain sacrifices conformational entropy for an increased potential energy by adopting a more ordered structure.

Although the radius of gyration plot gives insight on the collapse transition, there is relatively little change over the region of interest below the transition at \( T^* = 0.105 \) as shown in figure 5.4. At \( T^* = 0.105 \), the chain undergoes a further decrease in the size where there is a

![Figure 5.3: A plot of internal energy for two consecutive runs of the 32-mer peptide with \( \sigma/l \) equal to 1.6 and \( \lambda \) equal to 1.03.](image-url)
Chapter 5. Results and Discussion: effect of model parameters

Figure 5.4: Radius of gyration in terms of its three individual components $R_1$, $R_2$ and $R_3$ for the 32-mer with $\sigma/l$ equal to 1.6 and $\lambda$ equal to 1.03.

much smaller gradual decrease in the largest component of the $R_g$ while the other two components approach each other. This temperature overlaps with the low-temperature peak in the $C_v$ plot shown in figure 5.2. Accordingly, this property provides little insight on the low-temperature conformational change. Other properties should therefore be exploited.

Figure 5.5 is a histogram for probabilities of the RMSD between all possible pairs of structures across the high temperature collapse transition. The RMSD is a measure of difference between different structures at a given temperature to keep track of conformational change (detailed description is provided in section 4.2.3). Across the high-temperature transition at $T^* = 0.26$, the histograms show little or no change in RMSD values; at temperature above the transition the histogram is almost the same as the one below it. The exact nature of the intermediate formed (disordered or ordered) cannot be discussed before looking at the RMSD change across the low-temperature transition.

Figure 5.6 is a plot of RMSD between all pairs of monomers across the low-temperature transition at $T^* = 0.105$. By decreasing the temperature across the transition from $T^* = 0.128$ to $T^* = 0.04$ there is a large shift in the RMSD histograms to values closer to zero. Moreover, the RMSD histogram across the high-temperature transition shown in figure 5.5 is almost the same as the RMSD of the high-temperature structure known to be an extended disordered
5.1. Effect of number of beads ‘N’

Figure 5.5: A plot of RMSD histograms for the 32-mer with $\sigma/l$ equal to 1.6 and $\lambda$ equal to 1.03 across the collapse transition.

state. There is an insignificant shift in RMSD across the transition at $T^* = 0.205$ relative to the large shift in RMSD across the low-temperature transition shown in figure 5.6. This gives an indication that the high-temperature transition is a collapse to disordered structure (disordered-disordered transition), and the low-temperature one is disordered-ordered transition. However the histograms still cover a large range of RMSD values going from 0.5 to 1.5, therefore more analysis is required to identify the reason for the wide range of configurations at the lowest temperature. We look at the contacts formation/breaking as a function of temperature across the transitions for more insight.

Contact map is an analysis tool that gives a better understanding of the formation/breaking of the interactions between monomers (detailed description in section 4.2.3). Figure 5.7 shows the contact maps at three different temperatures: at $T^* = 0.461$ (above the high-temperature transition), $T^* = 0.21$ (across the high-temperature transition), and $T^* = 0.144$ (below the high-temperature transition). At a temperature above the transition, all the squares in the map are black, except for the white diagonal squares. This indicates that
the only contact formed is between monomers separated by a single monomer. This confirms that at high temperatures residues do not come within the range of interactions, and the model-peptide is an extended chain. Across the high-temperature transition and going to lower temperatures, the contact map in figure 5.7 (c) has two distinctive features; the appearance of white squares in the mid-section of the map, and the appearance of gray squares on the edges of the map. The white squares refer to stable contacts that persist through the time of the simulation while gray squares indicate that the contacts formed are sometimes broken. Accordingly, the contacts are of two types, permanent or transient. The disordered nature of the intermediate is reflected by a large number of transient contacts.

Figure 5.8 shows the contact maps at $T^* = 0.128$ (above the freezing transition), $T^* = 0.105$ (across the freezing transition), and $T^* = 0.04$ (below the freezing transition). Comparing the contact map of the first intermediate in figure 5.7 (c) at $T^* = 0.144$ with the contact map of the intermediate formed below the low-temperature transition at $T^* = 0.04$ in figure 5.8 (c), shows that very few contacts are formed at high temperature compared to those formed at low temperature. At $T^* = 0.04$, the contact map has relatively much less black squares than
5.1. **Effect of number of beads \( N \)**

**Figure 5.7**: Average contact maps of a 32-mer with \( \sigma/l \) equal to 1.6 and \( \lambda \) equal to 1.03 across the collapse transition. (a) Above the collapse transition, (b) Across the collapse transition, and (c) Below the collapse transition.

that at \( T^* = 0.144 \) and contains more white versus gray squares reflecting the permanent nature of the contacts when compared against the contact map at \( T^* = 0.144 \). Therefore the low-temperature intermediate is a compact ordered structure and the low-temperature transition is a disordered-ordered transition.

Figure 5.9 contains the RMSD histogram of the root mean square deviation at the lowest temperature. This plot can be used to explain the formation of an ordered structure after the low-temperature transition. The plot shows a bimodal distribution of structures: a main cluster at a value of 0.6, and a very small cluster with an RMSD of 1.7. It is not clear whether or not an RMSD deviation of 1.5 reflects structures with different contact maps as opposed to structures with different breathing motions associated with bond-length fluctuations.

The RMSD\(_{CM}\) histogram plot in figure 5.10 is presented to compare structures at a given temperature which can be used to cluster the structures according to variation in contacts rather than conformation. The RMSD\(_{CM}\) gives insight on the fraction of non-covalent bonds that are conserved across structures. Although, the RMSD histograms in conformation have been shown in figure 5.9, this does not mean that structures have the same contacts. Two
structures of equal RMSD will have a similar conformation but might have different contacts and vice versa. If there is conservation of contacts between structures (as it is expected in the ground state), the histogram is a tall narrow peak which approaches one in the limit where all the configurations have the same contact maps. Instead, the plot shows a bimodal distribution which indicates the presence of more than one cluster, with one major peak at a value of 0.9 and a smaller peak at a value of 0.3. This bimodal distribution is consistent with that found in figure 5.9 with similar areas in the peaks, which might indicate the pairs of configurations contributing to a large RMSD are similar to those contributing to a small RMSD$_{CM}$, in which case, changes in spatial configurations are reflected by changes in contact maps. This shows that there is a significant population of structures forming at least two distinct clusters of ordered structures that have different contact maps represented by the peak at low RMSD$_{CM}$ and as such non-native like behavior.

Accordingly, the bimodal in the RMSD$_{CM}$ with a main peak close to one coupled with the large shift in the RMSD towards lower value and the exchange of gray squares with white squares across the low-temperature transition, as shown in contact maps 5.8, indicates that the low-temperature transition is a disordered-ordered transition.
5.1. *Effect of number of beads ‘N’*

We look at the behavior of the core of the chain as a function of temperature to explain the different transitions and low-temperature fluctuations. The core is a group of residues (usually highly hydrophobic) buried inside the chain, inaccessible to a theoretical solvent, collectively known as the ‘core’, the ‘folding core’ or the ‘nucleus’ (Dinner et al., 2000). The solvent accessible surface area, or SASA analysis, provides the surface area of each bead that is exposed to a theoretical solvent molecule. A residue with a low SASA indicates that it is buried inside the chain where it cannot be accessed by a solvent. On the other hand, a high SASA value indicates that the residue is on the surface of the chain and therefore well-exposed to the solvent. Figure 5.11 shows the average SASA of each residue of the 32-mer chain at the lowest temperature. We define the core of the chain as the sum of residues with an average SASA of zero. The core is formed of 6 residues, 4 hydrophobic ones (valine, leucine, tryptophan and phenylalanine) and 2 polar residues (cysteine and tyrosine).

Figure 5.12 show the RMSC, which corresponds to the RMSD$_{CM}$ for only the spheres making up the core. This property might explain why there is a binodal distribution in the RMSD plots and large configurational fluctuations at low temperature. It is clear that at the lowest

![Figure 5.9: A plot of RMSD histograms for the 32mer with \( \sigma/l \) equal to 1.6 and \( \lambda \) equal to 1.03 at the lowest temperature.](image)
temperature, contacts of the core are not conserved as shown by the appearance of multiple peaks ranging from 0 to 1. This is consistent with the contact map at the lowest temperature as shown in figure 5.8 (c): the presence of some gray areas reflects the non-permanent contacts formed between the core groups and between the core and surface spheres. This is also not a characteristic of real proteins.

The SASA analysis is repeated for various temperatures across the two transitions to correlate the formation of the core with the folding behavior. Results are shown in figure 5.13 where the average SASA has been calculated for the chain at $T^*=0.04$. The residues forming the core are shown in red.

Figure 5.13 shows that at $T^*=0.26$ the first residue gets buried inside the chain. This corresponds to the high-temperature transition in $C_v$ and the temperature of collapse of the chain size and the beginning of formation of monomer-monomer interaction. Across the first transition at $T^*=0.144$, 4 residues are buried in the first intermediate. By comparing this chain with that at the lowest temperature, 3 out of the final 6 residues of the core are formed, while

---

**Figure 5.10:** A plot of RMSD histograms of the contact maps of the 32-mer chain with $\sigma/l$ equal to 1.6 and $\lambda$ equal to 1.03 at the lowest temperature.
5.1. **Effect of number of beads ’N’**

<table>
<thead>
<tr>
<th>Residue</th>
<th>SASA</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>18</td>
</tr>
<tr>
<td>T</td>
<td>13</td>
</tr>
<tr>
<td>S</td>
<td>19</td>
</tr>
<tr>
<td>T</td>
<td>22</td>
</tr>
<tr>
<td>N</td>
<td>18</td>
</tr>
<tr>
<td>H</td>
<td>14</td>
</tr>
<tr>
<td>M</td>
<td>14</td>
</tr>
<tr>
<td>I</td>
<td>12</td>
</tr>
<tr>
<td>Q</td>
<td>21</td>
</tr>
<tr>
<td>V</td>
<td>0</td>
</tr>
<tr>
<td>L</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
</tr>
<tr>
<td>I</td>
<td>7</td>
</tr>
<tr>
<td>P</td>
<td>10</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
</tr>
<tr>
<td>Y</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>6</td>
</tr>
<tr>
<td>L</td>
<td>8</td>
</tr>
<tr>
<td>D</td>
<td>17</td>
</tr>
<tr>
<td>A</td>
<td>10</td>
</tr>
<tr>
<td>G</td>
<td>16</td>
</tr>
<tr>
<td>D</td>
<td>16</td>
</tr>
<tr>
<td>A</td>
<td>10</td>
</tr>
<tr>
<td>G</td>
<td>15</td>
</tr>
<tr>
<td>K</td>
<td>11</td>
</tr>
<tr>
<td>E</td>
<td>22</td>
</tr>
<tr>
<td>S</td>
<td>8</td>
</tr>
<tr>
<td>G</td>
<td>8</td>
</tr>
<tr>
<td>R</td>
<td>16</td>
</tr>
<tr>
<td>V</td>
<td>20</td>
</tr>
<tr>
<td>K</td>
<td>14</td>
</tr>
<tr>
<td>W</td>
<td>0</td>
</tr>
</tbody>
</table>

**Figure 5.11:** Average SASA of each residue for the 32-mer chain at the lowest temperature. Red residues indicate core residues with SASA equal to zero.

one core buried at high temperature becomes exposed at the lowest temperature. Across the second transition 2 residue are buried and the lowest temperature has a core of 6 residues in total.

In fact the evolution of the core, along with previous analysis, shows that the high temperature transition is the collapse transition where the extended state collapses into a disorganized globule with 4 buried residues. We previously mentioned that the contact map of this intermediate in figure 5.8 (c) shows two types of contacts permanent and transient corresponding to two type of monomers. The SASA analysis showed that 67% of the core of the chain got buried and the remaining residues are exposed on the surface. The large exposed surface area coupled with the high temperature might explain the disordered state of the intermediate and wide range of RMSD histograms shown in figure 5.5. By further lowering the temperature, two residues are buried in the chain and form the core before undergoing the low-temperature transition. Across this transition, the clear contact map of white and
gray squares, along with the binodal clusters in RMSD and RMSD_{CM} with a main cluster in the limit of 1 show that the low-temperature transition is a freezing transition and the intermediate formed is an ordered intermediate. However, the binodal distribution at the lowest temperature suggests the existence of more than a cluster of ordered structures and a non-protein like behavior.

There are a number of reasons for observing more than one cluster of structures at low temperature. As shown in figure 5.11, 6 residues which are equivalent to 18% of the residues in the 32-mer chains are buried in the structures formed at the lowest temperature. This is a relatively small core size compared to real protein chains. Duan and Kollman (Duan and Kollman, 1998) reported 50% burial of residues in a 36 HP model of villin headpiece subdomain while Dobson and coworkers (Dobson, Sali, and Karplus, 1998) showed that for a 125-mer model 49% of the residues formed the core. Moreover, at the lowest temperature, the fact that a very small portion of the peptide is buried (18%) implies that most of the exposed surface is freely moving and contacts are not stabilized. If one or two contacts
5.1. Effect of number of beads ‘N’

in the core break, this is equivalent to saying that almost 33% of the contacts are not conserved. This might explain the multiple peaks in the histograms of the RMSC in figure 5.12 corresponding to a binodal RMSD$_{CM}$ and conformational fluctuations shown in the RMSD. Finally, 33% of the buried core consists of polar residues. This might explain the instability of the core as opposed to real protein. The expected folding of the 32-mer is that of small proteins: Taylor and coworkers (Taylor, Paul, and Binder, 2009a) argue that a large number of small proteins folds in a single-step process with a distinctive all-or-none characteristic, in which the unfolded extended peptide collapses directly into its ground state by the formation of the hydrophobic core. Instead the 32-mer undergoes 2 transitions to form a relatively unstable native state. This might be explained by the formation of a small core with 33% polar residues.
5.1.1.3 Folding transitions

In summary, the 32-mer model folding behavior has been investigated in detail in this section. It was found that the extended coil at high temperature undergoes a discontinuous transition, or a collapse transition at $T^* = 0.26$, into a disordered compact globule, where 67% of the low-temperature core residues are buried. The large change in energetic (internal energy) and conformational ($R_g$, conformation and contacts) properties takes place across the first transition, specifically the collapse of the chain. Below the collapse transition, the chain sacrifices conformational entropy for an increased potential energy by adopting a more compact structure. It has been argued that the collapse transition in real proteins is driven by the burial of this hydrophobic group of very strong contacts or the core. Dinner et al. (Dinner et al., 2000) postulated that the collapse can lead to a disorganized globule without specific contacts and with a relatively large configurational entropy. The configurational entropy of the collapsed globule is sufficiently small that its destabilizing contribution to the free energy on collapse is compensated by the burial of hydrophobic groups, even in the absence of native contacts. This was shown in the high-temperature transition and the first disordered compact intermediate. In fact, the appearance of few white squares in the contact maps across the high temperature transition can be identified as the contacts formed within the residues of the core, since most contacts in the chain are transient as shown with the mostly black and gray map. The intensity of the color indicates that the contacts only occur part time. This is consistent with what was predicted before: a hydrophobic core buried inside the peptide is formed across the transition, but the exposed surface of the chain, which have not established strong inter-monomer interactions, can move freely and sample a large number of configurations. Moreover, the second transition in the folding of the protein is identified as the freezing transition, where the previously collapsed structure orders itself to form well-defined contacts. The surface residues appear to move less by forming stronger bonds, as if they were ‘frozen’ in place, hence the name ‘freezing’ transition. This is a disordered-ordered transition as opposed to the high temperature disordered-disordered transition. At the lowest temperature, the conformation of the structure (figure 5.9), the contacts of the structure (figure 5.10) and the contacts in the core (figure 5.12) all fluctuate and therefore the gray shades that persist in the low-temperature contact map in figure 5.8 (c) might correspond to the ‘breathing motion’ of the proteins in their native state or more
probably a transitioning between two different conformational states reflecting a binodal in RMSD at the lowest temperature.

5.1.1.4 Short Discussion

The size of the core not only explains the low-temperature fluctuations, but also the unusual folding pathway. In fact, the effect of the exposure of core due to small perturbations on proteins fluctuations, unfolding and aggregation has been substantially explored in literature (Cromwell, Hilario, and Jacobson, 2006; Dinner et al., 2000; Dobson, Sali, and Karplus, 1998). The 32 monomer chain is expected to give results applicable to small proteins (Dobson, Sali, and Karplus, 1998). From the data analyzed, the 32-mer folds according to the following pathway: the disordered coil undergoes a collapse into a disordered globule with the burial of a hydrophobic core. This globule undergoes a continuous freezing transition into the native state, supposedly. However this folding pathway has not been reported before in literature. Small proteins with sufficiently short range interactions have a one state folding, where the two transitions collapse into one, and the extended coil folds into the native state in a single transition. The one stage folding is mainly driven by hydrophobic collapse of the core: proteins ball up into well-packed folded states in which the hydrophobic amino-acids are predominantly located in the protein’s core and the polar amino-acids are more commonly on the folded protein’s surface (Rampf, Binder, and Paul, 2006). Since the collapse of the hydrophobic core is the driving force of the folding of small proteins, the 32-mer did not behave similarly to small proteins because the loss in entropy could not be compensated by the formation of very few interactions with the small number of buried residues.

Accordingly, it can be concluded that the 32-mer is not a good protein model since neither the generic folding behavior nor the stability of the native conformation are displayed. This is consistent with the finding of Bratko et al. (Bratko et al., 2006) who argue that the 64-mer is the shortest chain capable of reproducing satisfactorily the folding behavior of proteins. In view of unrealistically low volume-to-area ratio, this short chain exhibits unrealistic folding behavior.
At this stage of the analysis, we are not looking at optimizing our system. Rather, we present an idea about our system, and the analysis tools employed. The length of the chain appears to have a major effect on the folding behavior, as a result, in the following section we investigate the folding behavior of a 64-mer model chain.
5.1. Effect of number of beads ‘N’

5.1.2 Folding behavior of a 64 monomer chain

The model-chain simulated is a 64 monomer chain with a long range of interaction potential $\lambda = 1.5$ and overlap parameter $\sigma/l = 1.9$. This particular sequence has been previously optimized for fast folding in isolation using on-lattice simulation (Bratko et al., 2006). However, the sequence has not been studied using an off-lattice square-well model. We use this sequence as a model system for demonstrating that the chain exhibits generic characteristics of protein folding. In this section we investigate the transitions and intermediates in the folding behavior, and identify the nature of the low-temperature structure with the ultimate aim of showing that the 64-mer model exhibits generic features of real proteins. We first introduce the model used, next we show the main results of the simulation and finally we conclude with a short discussion.

5.1.2.1 Model

In this section, we simulate an off-lattice 64-mer model-chain with residues interacting via the MJ potential. Simulation have been performed using replica exchange. Several runs using the same sequence with different starting configurations are performed in order to determine whether or not equilibration is reached. A detailed description of the model, interaction potential, and analysis tools common to all the simulations have been introduced in chapter 4. Here we describe the key input parameters and specific simulation details related to this section.

- **Sequence**: KEKSTAGRVASGVLDVSACGVLGDIDTLQGSPIALKTFYGNKFNDVE ASQAHMIRWPNYTLPE
- **Other parameters**: In this section, we use a range of interaction potential $\lambda$ equal to 1.5, which is the longest range of interaction used in this project, and an overlap parameter $\sigma/l$ equal to 1.9.
- **Equilibration run**: The system was equilibrated after 30 million time-steps to the lowest temperature.
- **Production run**: Two independent runs of 18 million time-steps each were used for data analysis.
- **Temperature list:** In table 5.2, we present the set of optimized temperatures of each replica in the simulation calculated as shown in sections 4.2.1 and 4.2.2. Temperatures are input in reduced units as described in section 4.1.3.

<table>
<thead>
<tr>
<th>List number</th>
<th>Temperature</th>
<th>List number</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.036</td>
<td>9</td>
<td>0.212</td>
</tr>
<tr>
<td>1</td>
<td>0.046</td>
<td>10</td>
<td>0.261</td>
</tr>
<tr>
<td>2</td>
<td>0.053</td>
<td>11</td>
<td>0.31</td>
</tr>
<tr>
<td>3</td>
<td>0.071</td>
<td>12</td>
<td>0.383</td>
</tr>
<tr>
<td>4</td>
<td>0.083</td>
<td>13</td>
<td>0.455</td>
</tr>
<tr>
<td>5</td>
<td>0.099</td>
<td>14</td>
<td>0.556</td>
</tr>
<tr>
<td>6</td>
<td>0.122</td>
<td>15</td>
<td>0.666</td>
</tr>
<tr>
<td>7</td>
<td>0.144</td>
<td>16</td>
<td>2.092</td>
</tr>
<tr>
<td>8</td>
<td>0.178</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 5.1.2.2 A general overview

The energy histograms of 17 optimized and reduced temperatures is displayed in figure 5.14. The temperature list is available in table 5.2.

The temperatures for the replica exchange are chosen such that there is sufficient overlap in energy histograms to obtain swapping efficiency in the order of 15 to 30%. The system was equilibrated to the lowest temperature, and the $C_v$ of 2 independent production runs of 18 million time-steps each overlap as shown in figure 5.15.

The specific heat capacity is a useful tool which gives insight on conformational change. The $C_v$ plot shown in figure 5.15 exhibits four well-defined peaks indicating that the chain undergoes four structural transitions at temperatures corresponding to the peaks in $C_v$. There is a high-temperature transition at $T^* = 0.556$, an intermediate temperature transition at $T^* = 0.212$, a second-lowest temperature transition at $T^* = 0.071$ and a low-temperature transition at $T^* = 0.046$. We refer to these transition as $T_1$, $T_2$, $T_3$ and $T_4$ respectively, and the average structures formed below these transitions as intermediates $I_1$, $I_2$, $I_3$, and $I_4$. The intermediate corresponds to the ensemble of conformations formed at a particular temperature. In particular cases where clustering is used, this intermediate can be further broken-down into smaller subsets or clusters, which also are referred to as intermediates. Every
transition and its intermediate are analyzed individually in order to better understand the folding behavior.

A simple tool to investigate monomer-monomer interaction is the internal energy. The reweighted internal energy is plotted as a function of temperature in figure 5.16. At high temperature, approximately between $0.556 < T^* < 2.2$, the sum of monomer-monomer interactions is constant and close to zero. Within this temperature range, the radius of gyration of the chain is also slightly decreasing from its maximum at 100 to 75 at $T^* = 0.556$ as shown in figure 5.17. Therefore it can be argued that the monomers are not sampling the square well potentials indicating that the chain is extended. At $T^* = 0.556$, the internal energy starts decreasing gradually, coupled with a sharp decrease in $R_g$ at the same temperature corresponding to the high-temperature peak in the heat capacity plot as shown in figure 5.15. This indicates that the chain is collapsing into a more compact structure driven by attractive square well interactions across the transition $T_1$, the collapse transition. At very low temperature, $T^* = 0.036$, the internal energy of the chain is lowest at around -75,
and the size of the chain is the smallest with an $R_g$ approximately equals to 13. This indicates that the residues are very close to each other forming multiple contacts and the chain therefore adopts a very compact structure.

The nature of the collapse transition, whether disordered-disordered or disordered-ordered, and accordingly the nature of the intermediate formed, needs more investigation and cannot be confirmed. Although the radius of gyration plot gives insight on the collapse transition, there is relatively little change in $R_g$ over the region below the transition which is of interest. We look at the change in contact maps, RMSD and RMSD$_{CM}$ for more information.

Figure 5.18 is a histogram plot of three RMSD histograms at temperatures $T^* = 0.261$ below the transition $T_1$ (equivalently above the transition $T_2$), $T^* = 0.212$ across the transition $T_2$, and $T^* = 0.144$ below the transition $T_2$. RMSD histogram at temperatures $0.556 < T^* < 2.2$ above and across the transition $T_1$ are not shown here as they almost overlap the histogram shown in black at $T^* = 0.261$ in figure 5.18. There is relatively no change between the RMSD at a temperature where the intermediate $I_1$ is formed when compared against

---

Figure 5.15: A plot of specific heat capacity of the 64-mer for two consecutive runs with $\sigma/l$ equal to 1.9 and $\lambda$ equal to 1.5.
high temperature where the peptide is in an extended state, on the other hand the RMSD histograms of the structures across the transition $T_2$ exhibit a large shift in the peak occurring at a value close to 1. Also the histogram of the intermediate formed at $T^* = 0.261$ is broad and samples a range of RMSD values while the histogram of the intermediate formed after the transition $T_2$ shown in green is a tall and narrow peak with an RMSD value slightly below 1 (around 0.8). This provides an indication that the high-temperature transition $T_1$ is a disordered-disordered transition while the transition $T_2$ is a disordered-ordered transition. These conformational fluctuations can be further explored by looking at how contact maps change with temperature.

Figure 5.19 contains the contact maps corresponding to temperatures occurring across the collapse transition $T_1$. At very high temperature ($T^* = 2.09$) the contact map (a) is predominantly black with no white squares indicating that no contacts between monomers are formed. This corresponds to a negligible internal energy and confirms that the chain is an extended random structure. The contact maps of structures at temperature across the collapse transition $T_1$ and at a temperature below it (figures 5.19 (b) and (c) respectively),
Figure 5.17: Radius of gyration in terms of its three individual components $R_1$, $R_2$ and $R_3$ for the 64-mer with $\sigma/l$ equal to 1.9 and $\lambda$ equal to 1.5.

Figure 5.18: A plot of RMSD histograms for the 64-mer with $\sigma/l$ equal to 1.9 and $\lambda$ equal to 1.5 across the second structural transition.
show that there are very few white squares with low intensity and the contact map is dominated by gray squares indicating that the peptide forms mainly transient contacts. This explains the deviations in conformations exhibited by a broad RMSD shown in figure 5.18 at \( T^* = 0.261 \). We can therefore conclude that the high-temperature transition \( T_1 \) is a collapse disordered-disordered transition. We next explore the nature of transitions \( T_2 \) and \( T_3 \) with their intermediates.

The contact map provides more insight on the nature of the transitions \( T_2 \) and \( T_3 \), and the structures formed. Figure 5.20 (a) is the contact map corresponding to \( T^* = 0.261 \) where intermediate \( I_1 \) is formed after the high-temperature transition, and contact map (b) is the contact map corresponding to the temperature equal to 0.144 below the transition \( T_2 \) where the intermediate \( I_2 \) is formed while contact map (e) is the contact map corresponding to \( T^* = 0.062 \) below the transition \( T_3 \) where the intermediate \( I_3 \) is formed. Contact maps (c) and (d) show the progression of the average contacts between the transitions \( T_2 \) and \( T_3 \). Comparing contact maps in figures 5.20 (a) and (b) of intermediates \( I_1 \) and \( I_2 \) shows that the intermediate \( I_2 \) with contacts map (b) has permanent contacts. The contact map (b) is formed of black and white squares only with negligible amount of gray shades while contact map (a) is predominantly gray with very few white squares. The appearance of clear white
squares reflects permanent monomer-monomer interactions and the formation of a well-defined configuration \( I_2 \). This correlates with the tall single narrow peak in RMSD shown in figure 5.18 at \( T^* = 0.212 \) around a value of 0.8 when compared against the broad histogram at higher value of 4 observed at temperature equal to 0.261. The large shift in RMSD to lower values coupled with well-defined contacts shown in the contact map confirms that the transition \( T_2 \) is a disordered-ordered transition where the disordered compact globule \( I_1 \) collapses into an ordered state \( I_2 \). This was rather unusual because the temperature is still relatively high compared to the lowest temperature of the simulation \( T^* = 0.036 \) across which the well-defined conformation is expected to form. Interestingly, contact map (e) of the intermediate \( I_3 \) is also formed of white and black squares only indicating that a well-defined structure is formed at this temperature. However, the location of the white squares on the map are different from those formed in contact map (b) and the contact maps do not overlap. This indicates that the intermediate \( I_3 \) is a structure with different contacts than the intermediate \( I_2 \). In fact, the formation of the intermediate \( I_3 \) can be tracked from the progression of contact maps (b) to (d): there is an exchange between many black and white squares, where many black squares in contact map (b) turn gray and then white in contacts maps (c) and (d) respectively, while white squares in contact map (b) fade in color in contact map (c) and turn black in contact map (d). This indicates that various preexisting contacts are broken while new contacts between previously non-interacting residues are established. Accordingly, going across the transition \( T_3 \), the peptide breaks old, yet well-defined bonds, and forms new well-defined bonds. This is clear by comparing contact maps (b) and (e): the two figures are representatives of two well-defined structures with white and black squares only, yet with a different location on the matrix.

Figure 5.21 is a histogram plot of RMSDs across the second-lowest temperature transition \( T_3 \) and shows the histogram of the intermediate \( I_3 \) at \( T^* = 0.062 \) which corresponds to the contact map 5.20 (e). It is clear that by reducing the temperature from \( T^* = 0.071 \) to \( T^* = 0.062 \) there is a shift in the RMSD histograms toward lower values: the RMSD histogram at \( T^* = 0.062 \) is around 0.5 whereas that of the histogram at \( T^* = 0.083 \) is around 0.7. Also, the peaks are slightly more defined, taller, and narrower. This shift is small compared to the shift across the transition \( T_2 \) shown in figure 5.18 from 3.5 to 1. This, coupled with the fact that contact maps of the intermediate \( I_2 \) and \( I_3 \) show well-defined contacts, indicates that
5.1. Effect of number of beads ‘\(N\)’

**Figure 5.20:** Average contact maps of the 64-mer with \(\sigma/l\) equal to 1.9 and \(\lambda\) equal to 1.5 across the second and third transition. (a) Above the second transition \((T^* = 0.261)\), (b) Below the intermediate transition \((T^* = 0.144)\), (c) Above the third transition \((T^* = 0.083)\), (d) Across the third transition \((T^* = 0.071)\), (e) Below the third transition \((T^* = 0.062)\).

The third transition \(T_3\) is an ordered-ordered transition.

**Figure 5.21:** A plot of RMSD histograms for the 64-mer with \(\sigma/l\) equal to 1.9 and \(\lambda\) equal to 1.5 across the second-lowest temperature transition.
We next investigate the nature of the well-defined contacts in intermediates $I_2$ and $I_3$ by comparing them to the contact map of the structure at the lowest temperature. The average contact map of the low-temperature structure at the lowest temperature is shown in figure 5.22. The contact map is almost entirely formed of black and white squares. The only difference between this contact map (at $T^* = 0.046$) and the contact map 5.20 (e) at $T^* = 0.062$ is the disappearance of most of the gray squares, without the appearance of any new white squares. This suggests that transition $T_4$ is actually an ordered-ordered transition. Across this transition, most of the native contacts are already permanent. However, there is a stabilization of the remaining transient contacts which correspond to surface groups, and hence confirming that the last transition corresponds to freezing these groups. By comparing this contact map to that of the intermediate $I_2$ in figure 5.20 (b) a lot of the white squares have similar location on the map indicating that many contacts formed in the intermediate $I_2$ are conserved to lower temperatures.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure522}
\caption{Average contact map of the 64-mer with $\sigma/l$ equal to 1.9 and $\lambda$ equal to 1.5 at the lowest temperature.}
\end{figure}
5.1. Effect of number of beads ‘N’

The RMSD histogram at the lowest temperature shown in figure 5.23 is a tall narrow peak at a value slightly above zero. This small shift in the RMSD could be related to the breathing motion of the chain at low temperatures which involves a slow collective movement and produces slight conformational change of the peptide. This explains the presence of trivial amount of gray squares in the contact map of the low-temperature structure in figure 5.22. This histogram indicates that a single cluster of peptides could be defined using a cut-off RMSD of 0.8. To this end, a clustering routine was used to find the representative structure of the system at the lowest temperature. 1500 structures were examined and 1490 of the 1500 were found to have an RMSD of 0.8 or more with one another. The representative structure of this cluster was taken as the native structure of the protein.

Another evidence of the formation of a unique native state is the stability of the core at low temperature. Using the SASA analysis, we define the residues with a low SASA (<5) as buried residues. The choice of cutoff is subjective and depends on the progression of the average SASA value of the residues across the temperature range. These are shown in the following section, but at the lowest temperature it was clear that the residues have an average SASA either less than 5 or more than 10. The average SASA of the 64-mer chain at the lowest temperature \( T^* = 0.036 \) are shown in figure 5.24. The residues in red are
hydrophobic residues including glycine (G), alanine (A), valine (V), leucine (L), isoleucine (I), proline (P), phenylalanine (F), methionine (M), and tryptophan (W). Polar residues on the other hand are labeled in blue and include threonine (T), histidine (H), and tyrosine (Y).

At the lowest temperature 29 residues are buried with 25 hydrophobic residues and 4 polar ones. Accordingly the core constitutes 45% of the chain, which is consistent with literature. Dobson and coworkers (Dobson, Sali, and Karplus, 1998) showed that for a 125-mer model 49% of the residues formed the core. The slightly smaller core might be due either to the longer range of interaction employed in this model, or because the 64-mer chain is simply smaller and has fewer residues in total. This makes sense as surface area to volume ratio goes with inverse of $N$ to the one third.

Figure 5.25 is a histogram of probability for contacts formed only between core groups at $T^* = 0.036$. The histogram is a tall, well-defined, narrow peak at RMSC equal to 1 indicating the existence of specific stable contacts between buried residues. As expected, a very small peak appears at an RMSC equal to 0.6 which can be identified with the breathing motion of the core at the lowest temperature. These transient contacts in the core can be assumed to drive contact fluctuations in the whole chain which is manifested by conformational fluctuations. This might explain the slight shift in RMSD in figure 5.23 and the appearance of negligible gray squares in the contact map of the low-energy conformation in figure 5.22. The low-temperature structure has a well-buried hydrophobic core (86% of the buried residues are hydrophobic ones) with well-defined contact and does not exhibit structural fluctuations. This confirms that this structure exhibits characteristics of native protein chains.

In summary, we established that the 64-mer chain folds after 4 transitions from a high-temperature extended structure to a low-temperature native state. The transition $T_1$ is a disordered-disordered transition where the extended chain collapses into a more compact disordered globule $I_1$. By lowering the temperature, the globule undergoes a disordered-ordered transition $T_2$ at which the intermediate $I_1$ arrange into a well-defined structure $I_2$ with mostly native contacts that persists to low-temperatures. The transition $T_3$ is an ordered-ordered transition where the intermediate breaks some of the well-defined contacts and forms new well-defined contacts in the intermediate $I_3$. This intermediate undergoes an ordered-ordered transition into the native low-temperature structure where some of the
5.1. Effect of number of beads ‘N’

<table>
<thead>
<tr>
<th>Residue</th>
<th>Average SASA</th>
<th>Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>60</td>
<td>4.6</td>
</tr>
<tr>
<td>E</td>
<td>18</td>
<td>4.1</td>
</tr>
<tr>
<td>K</td>
<td>16</td>
<td>4.6</td>
</tr>
<tr>
<td>S</td>
<td>13</td>
<td>4.9</td>
</tr>
<tr>
<td>T</td>
<td>4</td>
<td>5.0</td>
</tr>
<tr>
<td>A</td>
<td>4</td>
<td>3.9</td>
</tr>
<tr>
<td>G</td>
<td>5</td>
<td>4.1</td>
</tr>
<tr>
<td>R</td>
<td>10</td>
<td>4.6</td>
</tr>
<tr>
<td>V</td>
<td>0</td>
<td>2.8</td>
</tr>
<tr>
<td>A</td>
<td>12</td>
<td>2.9</td>
</tr>
<tr>
<td>S</td>
<td>17</td>
<td>3.6</td>
</tr>
<tr>
<td>G</td>
<td>4</td>
<td>2.7</td>
</tr>
<tr>
<td>V</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>L</td>
<td>0</td>
<td>1.1</td>
</tr>
<tr>
<td>D</td>
<td>13</td>
<td>4.1</td>
</tr>
<tr>
<td>S</td>
<td>22</td>
<td>3.2</td>
</tr>
<tr>
<td>V</td>
<td>5</td>
<td>3.8</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>2.9</td>
</tr>
<tr>
<td>C</td>
<td>10</td>
<td>1.0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>V</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>L</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0.3</td>
</tr>
<tr>
<td>D</td>
<td>15</td>
<td>3.9</td>
</tr>
<tr>
<td>I</td>
<td>11</td>
<td>2.9</td>
</tr>
<tr>
<td>D</td>
<td>20</td>
<td>3.4</td>
</tr>
<tr>
<td>T</td>
<td>10</td>
<td>2.9</td>
</tr>
<tr>
<td>L</td>
<td>0</td>
<td>0.4</td>
</tr>
<tr>
<td>Q</td>
<td>12</td>
<td>3.3</td>
</tr>
<tr>
<td>G</td>
<td>11</td>
<td>3.6</td>
</tr>
<tr>
<td>S</td>
<td>11</td>
<td>3.0</td>
</tr>
<tr>
<td>P</td>
<td>12</td>
<td>3.1</td>
</tr>
<tr>
<td>I</td>
<td>0</td>
<td>1.1</td>
</tr>
<tr>
<td>A</td>
<td>4</td>
<td>3.1</td>
</tr>
<tr>
<td>K</td>
<td>14</td>
<td>2.6</td>
</tr>
<tr>
<td>L</td>
<td>4</td>
<td>2.1</td>
</tr>
<tr>
<td>K</td>
<td>17</td>
<td>2.3</td>
</tr>
<tr>
<td>T</td>
<td>11</td>
<td>1.1</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>Y</td>
<td>10</td>
<td>0.0</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1.0</td>
</tr>
<tr>
<td>N</td>
<td>18</td>
<td>2.9</td>
</tr>
<tr>
<td>K</td>
<td>23</td>
<td>2.7</td>
</tr>
<tr>
<td>F</td>
<td>12</td>
<td>2.6</td>
</tr>
<tr>
<td>N</td>
<td>17</td>
<td>3.9</td>
</tr>
<tr>
<td>D</td>
<td>11</td>
<td>3.8</td>
</tr>
<tr>
<td>V</td>
<td>0</td>
<td>2.1</td>
</tr>
<tr>
<td>E</td>
<td>18</td>
<td>4.5</td>
</tr>
<tr>
<td>A</td>
<td>4</td>
<td>4.0</td>
</tr>
<tr>
<td>S</td>
<td>23</td>
<td>3.4</td>
</tr>
<tr>
<td>Q</td>
<td>23</td>
<td>2.6</td>
</tr>
<tr>
<td>A</td>
<td>12</td>
<td>2.3</td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>0.4</td>
</tr>
<tr>
<td>M</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>I</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>R</td>
<td>10</td>
<td>0.3</td>
</tr>
<tr>
<td>W</td>
<td>2</td>
<td>1.2</td>
</tr>
<tr>
<td>P</td>
<td>12</td>
<td>2.7</td>
</tr>
<tr>
<td>N</td>
<td>13</td>
<td>3.4</td>
</tr>
<tr>
<td>Y</td>
<td>3</td>
<td>1.6</td>
</tr>
<tr>
<td>T</td>
<td>3</td>
<td>2.2</td>
</tr>
<tr>
<td>L</td>
<td>0</td>
<td>0.9</td>
</tr>
<tr>
<td>P</td>
<td>1</td>
<td>3.8</td>
</tr>
<tr>
<td>E</td>
<td>35</td>
<td>5.2</td>
</tr>
</tbody>
</table>

**Figure 5.24:** Average SASA of the 64-mer residues at the lowest temperature $T^* = 0.036$, with error deviation. Residues in red and blue are core hydrophobic and polar residues respectively.
transient contacts either become well-established or do not occur anymore. We next try to explain this folding behavior and discuss some of these intermediates and transitions by relating them to other work done in the same field.

### 5.1.2.3 Folding transitions and intermediates

After identifying the major properties of the transitions and intermediates in the folding behavior of the 64-mer, we show in table 5.3 the formation of the core as a function of temperature across all transitions for more insight. The average SASA of the residues is calculated for structures at the 17 temperatures shown in table 5.2. The cutoff of 8 is chosen by examining the change in average SASA with temperature; there is a high shift in SASA from the value of 8 or above to very low values, approximately equal to zero. In table 5.3 we report the core residues of the structures at temperature corresponding to transitions and intermediates studied in the previous section; core residues are colored in red and blue, red for hydrophobic residues and blue denoting polar residues. At the lowest temperature, the native structure has 29 buried residues, with 25 being hydrophobic ones.
We compare the core of a structure at a given temperature with that of the native structure: across the transition \( T_1 \) at \( T^* = 0.556 \), 15 residues are buried, 14 being hydrophobic and one polar. Out of the 15 residues, 13 hydrophobic are preserved in the low-temperature structure which indicates that the high-temperature transition is a hydrophobic collapse transition where 45% of the native core is formed in a single step. Accordingly the appearance of white squares in the contact map of the structure at \( T^* = 0.556 \) shown in figure 5.19 (b) can be explained by strong contacts formed between the hydrophobic core residues whereas the dominance of gray shades implies that the contacts only occurring part of the time can be related to the exposed surface residues of the chain which constitute 75% of the total residues. These surface residues can move freely and sample a large number of configurations. Contact map of the structure at \( T^* = 0.455 \) in figure 5.19 (c) have slightly more white squares than that of structure at \( T^* = 0.556 \) and has in fact one more residue buried as shown in table 5.3. This provides an indication that the appearance of white squares can be correlated with contacts between core residues.

Next we look at the transition \( T_2 \) and intermediate \( I_2 \). The transition \( T_2 \) at \( T^* = 0.212 \) was identified as a disordered-ordered transition where the collapse globule arranges into a more ordered intermediate \( I_2 \) at \( T^* = 0.144 \). The contact map of the intermediate \( I_2 \) in figure 5.20 (b) is formed of white and black squares only, which indicates a well defined structure. However, certain contacts were not conserved by lowering the temperature as contact maps in figure 5.20 (b) and (e) of the intermediates \( I_2 \) and \( I_3 \) respectively do not

**Table 5.3:** SASA analysis for the 64-mer across the temperature range with a cutoff = 8. Hydrophobic residues are labeled in red while polar residues are labeled in blue.

<table>
<thead>
<tr>
<th>( T ) Chain</th>
<th>KEKSTAGRVASGVLDSVACGVLGEDTQLGSPIAK1KTFYGNKNDVEASQAHIWPMYTLPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.036</td>
<td>KEKSTAGRVASGVLDSVACGVLGEDTQLGSPIAK1KTFYGNKNDVEASQAHIWPMYTLPE</td>
</tr>
<tr>
<td>0.046</td>
<td>KEKSTAGRVASGVLDSVACGVLGEDTQLGSPIAK1KTFYGNKNDVEASQAHIWPMYTLPE</td>
</tr>
<tr>
<td>0.062</td>
<td>KEKSTAGRVASGVLDSVACGVLGEDTQLGSPIAK1KTFYGNKNDVEASQAHIWPMYTLPE</td>
</tr>
<tr>
<td>0.071</td>
<td>KEKSTAGRVASGVLDSVACGVLGEDTQLGSPIAK1KTFYGNKNDVEASQAHIWPMYTLPE</td>
</tr>
<tr>
<td>0.083</td>
<td>KEKSTAGRVASGVLDSVACGVLGEDTQLGSPIAK1KTFYGNKNDVEASQAHIWPMYTLPE</td>
</tr>
<tr>
<td>0.144</td>
<td>KEKSTAGRVASGVLDSVACGVLGEDTQLGSPIAK1KTFYGNKNDVEASQAHIWPMYTLPE</td>
</tr>
<tr>
<td>0.212</td>
<td>KEKSTAGRVASGVLDSVACGVLGEDTQLGSPIAK1KTFYGNKNDVEASQAHIWPMYTLPE</td>
</tr>
<tr>
<td>0.261</td>
<td>KEKSTAGRVASGVLDSVACGVLGEDTQLGSPIAK1KTFYGNKNDVEASQAHIWPMYTLPE</td>
</tr>
<tr>
<td>0.455</td>
<td>KEKSTAGRVASGVLDSVACGVLGEDTQLGSPIAK1KTFYGNKNDVEASQAHIWPMYTLPE</td>
</tr>
<tr>
<td>0.556</td>
<td>KEKSTAGRVASGVLDSVACGVLGEDTQLGSPIAK1KTFYGNKNDVEASQAHIWPMYTLPE</td>
</tr>
</tbody>
</table>
overlap ($I_3$ being the intermediate formed at $T^* = 0.062$ across the transition $T_3$). Comparing the core of the structures with each other and with that of the native structure might give an indication about native and non-native contacts. The intermediate $I_2$ has 31 residues in total: 23 hydrophobic and 8 polar residues, where 3 hydrophobic residues in the native core do not appear in the core of $I_2$, and one additional hydrophobic residue buried in $I_2$ is not present in the native core. Moreover, 5 polar residues buried in $I_2$ do not take part of the core of the native state, while one polar residue appearing in the native core is on the surface of the chain $I_2$. Accordingly, there is an 86% conservation in total core residues.

Moreover, the rather unusual appearance of a well-defined structure at high temperature can be explained by the formation of a large core in intermediate $I_2$ even larger than that of the native state. If the stability of the native state of a chain is driven by the formation of certain favorable contact between buried groups, then the 82% similarity between the core of the intermediate $I_2$ and the native state might provide some explanation to the presence of such well-defined contacts and the native-like characteristics of this intermediate.

Figure 5.26 is the histogram plot of the RMSD of contacts of the core, or RMSC, across the transition $T_2$ relative to the contacts of the core of the native state. After the collapse transition, at temperature equal to 0.261, some of the core is forming as shown by the SASA analysis. This is reflected by a broad RMSD at high values. The red line represents the RMSC histogram of the core of the intermediate $I_2$ at $T^* = 0.144$. Two peaks appear with a predominant tall and narrow peak shifted toward a value of 1, and a small broad peak around 0.7. The presence of a tall narrow peak close to 1 provides an indication that the core of the intermediate $I_2$ has native-like contacts related to the hydrophobic residues buried. On the other hand, non-native contacts are mostly related to a change in the burial or exposure of polar residues. These appear to suffer the most change with temperature and are therefore the least stable.

It has been argued that in one of the folding behavior of proteins, the extended coil collapses into a compact disordered globule. This disorganized globule undergoes a disorder-order transition into an intermediate with native-like secondary structure, which is referred to as the molten globule (Karplus et al., 1995). This intermediate has highly probable contacts and a lower configurational entropy. The molten globule structure has been extensively studied in both experiments and simulations (Bam et al., 1998; Semisotnov et al., 1996;
5.1. Effect of number of beads ‘N’

Hammarstrom et al., 1999). It is a structure found in the folding process of real proteins with three basic distinct characteristics. First, they have globular compact nature, secondly they are well-defined stable structures, and finally they contain native secondary structures (Bam, Cleland, and Randolph, 1996; Dinner et al., 2000). The intermediate shown in figure 5.20 (b) is a well defined native-like structure which occurs after two transitions, but is not exactly identical to the low-temperature structure with contacts shown in figure 5.20 (e) or 5.22. Moreover, it has been argued that for the molten globule to occur, the hydrophobic core should start forming early in the folding process (Dinner et al., 2000). The collapse transition of the 64-mer chain occurs at $T^* = 0.556$ and that of the 32-mer model occurs at $T^* = 0.26$. Accordingly, the core of the 64-mer chain starts forming much earlier than that of the 32-mer chain. Dinner et al. (Dinner et al., 2000) argue that a free-energy preference for the burial of hydrophobic core residues while keeping polar beads exposed to the solvent on the surface of the collapsed chain drives the chain into native-like regions of conformational space, which is the molten globule. In short, the intermediate transition is a disorder-order transition, and the second intermediate $I_2$ is the molten globule exhibiting all expected characteristics.

The nature of the molten globule is confirmed by looking at lower-temperature structures.
and comparing them to the native state. Structures formed at \( T^* = 0.083 \), \( T^* = 0.071 \) and \( T^* = 0.062 \) show the progression from the molten globule to the low-energy native structure across the transition \( T_3 \). The contacts formed in these structures appear in contact maps 5.20 (c), (d) (structure at transition \( T_3 \)), and (e) (structure of intermediate \( I_3 \)), and the progression of the core in these structure is shown in table 5.3. First, the transition \( I_3 \) was established to be an ordered-ordered transition. Secondly, the intermediate \( I_3 \) appears to have 28 buried residue with only one fewer polar residue present in the core of the native structure with a total of 97% similarity. The similarity with the native state increases from 80% at \( T^* = 0.122 \), to 85% at \( T^* = 0.083 \), and finally to 97% at \( T^* = 0.071 \) and \( T^* = 0.062 \). There is only a change in 2 hydrophobic residues while there is a exchange in 5 polar residues in total across this temperature range. Accordingly, the transition \( T_3 \) is an ordered-ordered transition where the molten globule undergoes a search for the native state by initially burying the core of the low-temperature structure. Once the core is formed the rest of the chain arranges itself into the native conformation. In fact the presence of the gray shades in contact map 5.20 (c) can be attributed to the fluctuations of the polar residue threonine which is yet to be buried.

A main validation of the model as an acceptable representation of a protein is the formation of the native state at the lowest temperature. The transition \( T_4 \) occurring at \( T^* = 0.046 \) was also identified as an ordered-ordered freezing transition where the intermediate \( I_3 \) folds into the native state. In fact the core of the structure at \( T^* = 0.046 \) is the same as the core at the lowest temperature as shown in table 5.3. By comparing contact maps of the structures below the transition and at the lowest-structures in figures 5.20 (e) and 5.22 respectively, it is clear that most of the gray squares in the contact map of the intermediate \( I_3 \) disappear in the contact map of the low-temperature structure, although both contact maps have the same white squares. This confirms even further that the transition \( T_4 \) is a freezing transition where the fluctuations of exposed residues are reduced and the structure at the lowest temperature is a native low-energy structure with all the characteristics of the native state: a well-defined structure and a buried core with specific interactions.

### 5.1.2.4 A short discussion

In summary, the folding behavior of the 64-mer was identified as follows: at high temperature the extended peptide undergoes first a disordered-disordered collapse transition
5.1. Effect of number of beads ‘N’

<table>
<thead>
<tr>
<th>32mer</th>
<th>PTSTHNMIQLCPYALDGDAKESGRVKW</th>
</tr>
</thead>
<tbody>
<tr>
<td>64mer</td>
<td>KEKSTAGRVASVLDSDYACGVLGDIIDTLQGSPIAKLKTYGKNDVEASQAHMIRPWNYTLP</td>
</tr>
</tbody>
</table>

**Figure 5.27:** A table showing the solvent accessible surface area of the 32-mer and the 64-mer peptides at the lowest temperature. Residues in red are the ones which constitute the core of the chain.

at $T^* = 0.556$ leading to the formation of few strong interaction and a compact disordered globule with a hydrophobic core $I_1$. This compact disordered structure undergoes a disordered-ordered transition through an intermediate transition at $T^* = 0.212$ leading to the formation of a well-defined compact structure with some native interaction, the molten globule $I_2$. Across the transition $T_3$ at $T^* = 0.071$, the molten-globule undergoes an ordered-ordered transition where the non-native bonds are replaced with native ones, forming a core identical to that of the native structure. However the structure is still fluctuating a bit. Finally, the transition $T_4$ is identified as the freezing transition where the intermediate $I_3$ freezes into the native state and the minor fluctuations of surface polar residues are frozen.

The main issue being addressed in this first section is to choose a proper length for the sequence by which the generic behavior of the protein folding is reproducible. It is always tempting to simulate shorter chains giving a low computational cost and an easy equilibration. However, many researchers investigated the shortest possible chain capable of reproducing satisfactorily the folding behavior of proteins. Both of the 32-mer and the 64-mer folding behavior were analyzed, and the analysis showed that the 32-mer is not a good model whereas the 64-mer is a promising one exhibiting all characteristics of protein folding.

Figure 5.27 compares the cores of the 32-mer and 64-mer model chains. The residues in red form the core of the molecule. The 32-mer sequence has a core formed of 6 residues. On the other hand, the 64-mer has a core of 29 residues. Therefore, if for example two contacts break in the core of the 32-mer chain, it is equivalent to saying that 33% of the contacts of the core are not conserved, and 6% of the contacts of the core of the 64-mer are not conserved. This implies that the 64-mer is much more stable than the 32-mer which exhibits an unrealistically small core-to-chain ratio, and is therefore a better model used in simulating protein folding and unfolding.

The main objective of this section is not to analyze the folding behavior of various protein
chains, rather we show that for a given set of parameters our model can reproduce satisfactorily the folding characteristics of real protein. We also introduced the popular analysis tools used in this project.

The main conclusions of this section are as follows:

- The 32-mer model folds in a two-step folding process: a collapse transition forming a compact disorganized structure with a hydrophobic core, and a freezing transition forming an unstable low-temperature structure.

- The 64-mer model folds in a four-step folding process: a collapse transition forming a compact disorganized structure with a hydrophobic core, a disordered-ordered transition forming a molten globule, an ordered-ordered transition yielding a well-defined native-like structure with slight fluctuations, and a freezing transition towards the native state where most of previous fluctuations are frozen-out.

- The 32-mer chain is not a good model for protein folding because the core, driving the folding process and holding the chain all-together, is unrealistically small.

- The 64-mer peptide appears as a good model which exhibits folding characteristics of proteins, and reproduces satisfactorily the generic behavior of real protein peptides.

Although the 64 monomer chain displayed all characteristics of real proteins, the folding behavior displayed multiple transitions for the set of parameters chosen. Accordingly, the following sections will examine the folding behavior of larger chains in order to choose the optimum size for our model.
5.1.3 Folding of longer chains

In this section we study the folding behavior of a 128-mer and a 256-mer chains with $\lambda = 1.3$ and $\sigma/l = 1.3$ in order to investigate the effect of chain length (or number of residues $N$) on protein folding. In the previous sections, we studied the behavior of the 32-mer and the 64-mer, however, the computational cost increased significantly for larger chains by requiring a lot of time to equilibrate the systems and analyze them. Since the ultimate aim of this work is to study protein aggregation using multichain systems, the choice of very long chain ($N \geq 128$) seemed inefficient.

Accordingly, in this section we present the main data from the simulations of 128-mer and 256-mer chains and show a general overview of the folding behavior to check whether it is consistent with the previous findings. We first introduce the model used, next we show the main results, and finally we conclude with a short discussion.

5.1.3.1 Model

In this section, we simulate off-lattice 128-mer and 256-mer model chains with residues interacting via the MJ potential. Simulation have been performed using replica exchange. Several runs using the same sequence with different starting configurations are performed in order to determine whether or not equilibration is reached. A detailed description of the models, interaction potential, and analysis tools common to all the simulations have been introduced in chapter 4. Here we describe the key input parameters and specific simulation details related to this section.

Details of the 128-mer model chain are listed below:

- **Sequence**: KEKSTAGRVASGVLDSVACGVLGDIIDLQGSPAISKLFYGNKFDV
  ASQAAMIRWPNYTLPEKEKSTAGRVASGVLDSVACGVLGDIIDLQGSPAISKLF
  YGNKFDVSEASQAAMIRWPNYTLPE

- **Other parameters**: The range of interaction potential $\lambda$ is set to 1.3, and the overlap parameter $\sigma/l$ is equal to 1.3.

- **Equilibration run**: The system was equilibrated after 100 million time-steps to the lowest temperature.
- **Production run**: Two independent runs of 30 million time-steps each were used for data analysis.

- **Temperature list**: In Table 5.4, we present the set of optimized temperatures of each replica in the simulation calculated as shown in Sections 4.2.1 and 4.2.2. Temperatures are input in reduced units as described in Section 4.1.3.

  **Table 5.4**: Simulation-temperatures list of the 128-mer and 256-mer.

<table>
<thead>
<tr>
<th>List number</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.040</td>
</tr>
<tr>
<td>1</td>
<td>0.046</td>
</tr>
<tr>
<td>2</td>
<td>0.053</td>
</tr>
<tr>
<td>3</td>
<td>0.071</td>
</tr>
<tr>
<td>4</td>
<td>0.083</td>
</tr>
<tr>
<td>5</td>
<td>0.105</td>
</tr>
<tr>
<td>6</td>
<td>0.128</td>
</tr>
<tr>
<td>7</td>
<td>0.144</td>
</tr>
<tr>
<td>8</td>
<td>0.178</td>
</tr>
</tbody>
</table>

  
<table>
<thead>
<tr>
<th>List number</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>0.212</td>
</tr>
<tr>
<td>10</td>
<td>0.261</td>
</tr>
<tr>
<td>11</td>
<td>0.31</td>
</tr>
<tr>
<td>12</td>
<td>0.383</td>
</tr>
<tr>
<td>13</td>
<td>0.455</td>
</tr>
<tr>
<td>14</td>
<td>0.556</td>
</tr>
<tr>
<td>15</td>
<td>0.666</td>
</tr>
<tr>
<td>16</td>
<td>2.092</td>
</tr>
</tbody>
</table>

For the 256-mer, the model is described as it follows:

- **Sequence**: KEKSTAGRVASGVLDIVERACGVLGDIDTLQGSPIAKLKTGFYGNKFNDVE ASQAHRWRPNYTLPEKEKSTAGRVASGVLDIVERACGVLGDIDTLQGSPIAKLKTGF YGNKFNDVESQAHRWRPNYTLPEKEKSTAGRVASGVLDIVERACGVLGDIDTLQ GSPIAKLKTGFYGNKFNDVESQAHRWRPNYTLPEKEKSTAGRVASGVLDIVERACG VLGDIDTLQGSPIAKLKTGFYGNKFNDVESQAHRWRPNYTLPE

- **Other parameters**: The range of interaction potential $\lambda$ is set to 1.3, and the overlap parameter $\sigma/l$ equal to 1.3.

- **Equilibration run**: The system was equilibrated after 180 million time-steps to the lowest temperature.

- **Production run**: Two independent runs of 30 million time-steps each were used for data analysis.

### 5.1.3.2 A general overview

Figure 5.28 is a plot of the specific heat capacity as a function of temperature for a 64-mer and a 128-mer model chains. In order to test the effect of chain length only, all other parameters
are held constant ($\sigma = 1.3$, and $\lambda = 1.3$). The 64-mer chain shows a dominant peaks at $T^* = 0.3$ and a predominant peak at $T^* = 0.12$, while the heat capacity plots of the 128-mer and the 256-mer exhibit a single peaks at temperatures equal to 0.2 and 0.12 respectively. Increasing the number of residues reduces the number of transitions in the folding of the chain, and shifts the transitions to lower temperatures.

![Figure 5.28: A plot of specific heat capacity as a function of temperatures for 3 chains with different lengths: 64-mer, 128-mer, and 256-mer. The three chains have a $\lambda$ equal to 1.3 and $\sigma$ equal to 1.3.](image)

Figure 5.29 is a plot of the radius of gyration for a 64-mer, a 128-mer and a 256-mer chains with $\lambda = 1.3$ and $\sigma = 1.3$ as a function of reduced temperature. The temperature at which the size of the chain starts decreasing corresponds to the temperature of the collapse transition. Here, the high-temperature peak at $T^* = 0.3$ of the 64-mer, and the single peaks for longer chains are identified as the collapse transitions.

To better characterize the nature of these collapse transitions (disordered-disordered, or ordered-ordered), and the characteristics of the low-temperature states, we look at contact maps, and the change in RMSD and RMSD$_{CM}$ with temperature. We then define the core
Chapter 5. Results and Discussion: effect of model parameters

Figure 5.29: A plot of the radius of gyration as a function of temperature for 3 chains with different lengths: 64-mer, 128-mer, and 256-mer. The three chains have a $\lambda$ equal to 1.3 and $\sigma$ equal to 1.3.

Based on the data found in Table 5.5, the 64-mer folds after two transitions from the extended state to a well-defined native state. These two transitions are identified as a high-temperature collapse transition (at $T^* = 0.4$) and a low-temperature freezing transition (at $T^* = 0.12$). Across the collapse transition, the random high-T state collapses into an ordered compact globule with native structure. Although 93% of the native residues are buried in this intermediate state, there is some contact- and conformational-fluctuations correlated to the appearance of some gray squares in the contact map at this temperature. Across the freezing transition, these fluctuations are frozen out and a low-temperature well-defined structure is formed. On the other hand, both of the 128-mer and the 256-mer chains undergo a direct transition (at $T^* = 0.2$ and $T^* = 0.12$ respectively) from the extended high-T state to a low-temperature well-defined structure with a large core. Accordingly, by increasing the number of residues in a chain, the collapse and the freezing transitions merge into a single transition with all-or-none characteristics.
**5.1. Effect of number of beads ‘N’**

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Property</th>
<th>64-mer</th>
<th>128-mer</th>
<th>256-mer</th>
</tr>
</thead>
<tbody>
<tr>
<td>T* = 0.1</td>
<td>1. at bottom of the first peak Min. Rg value</td>
<td>1. at the peak Slightly above the min value</td>
<td>Bottoms of the peak size starts decreasing</td>
<td>1. at the peak Slightly above the min value</td>
</tr>
<tr>
<td>Contact maps</td>
<td>Single tall peak around 1.5</td>
<td>Single tall peak around 3</td>
<td>Black square, very few white squares</td>
<td>Single tall peak around 3</td>
</tr>
<tr>
<td>RMSD</td>
<td>Single tall peak around 0.8</td>
<td>Single tall peak around 0.8</td>
<td>Black and white with multiple grey squares</td>
<td>Single tall peak around 0.8</td>
</tr>
<tr>
<td>Number of core residues (</td>
<td></td>
<td></td>
<td>Black and white with multiple grey squares</td>
<td>Black and white with multiple grey squares</td>
</tr>
<tr>
<td>Transition Type</td>
<td>Collapse transition</td>
<td>Multiple peaks</td>
<td>Slightly above the min value</td>
<td>Multiple peaks</td>
</tr>
<tr>
<td>State Characteristics</td>
<td>disordered-ordered trans.</td>
<td>Multiple peaks</td>
<td>Local structure, cannot be defined yet</td>
<td>N/A</td>
</tr>
<tr>
<td>T* = 0.2</td>
<td>1. at bottom of the second transition</td>
<td>1. at the transition</td>
<td>Collapse transition</td>
<td>1. at the transition</td>
</tr>
<tr>
<td>Contact maps</td>
<td>Single tall peak at 1</td>
<td>Single tall peak around 0.8</td>
<td>Single tall peak around 0.8</td>
<td>Single tall peak around 0.8</td>
</tr>
<tr>
<td>RMSD</td>
<td>Single tall peak at 0</td>
<td>Single tall peak at 0</td>
<td>N/A</td>
<td>Single tall peak at 0</td>
</tr>
<tr>
<td>Number of core residues (</td>
<td></td>
<td>30, (100%)</td>
<td>44, (150%)</td>
<td>44, (150%)</td>
</tr>
<tr>
<td>Transition Type</td>
<td>ordered-ordered freezing trans.</td>
<td>disordered-ordered trans.</td>
<td>From extended to native state</td>
<td>From extended to native state</td>
</tr>
<tr>
<td>State Characteristics</td>
<td>native state, well defined struct.</td>
<td>native state, well defined struct.</td>
<td>native state, well defined struct.</td>
<td>native state, well defined struct.</td>
</tr>
</tbody>
</table>

**5.1.3.3 A short discussion**

Many folding patterns have been detected so far, including a single-step transition, two-step folding, and multiple-step folding. In fact, one of the main problems that seems ubiquitous across the wide range of biological systems is the folding process in which the protein undergoes a symmetry-breaking transition from an expanded random state to a compact ordered rigid state. We attributed the main difference between the folding of the 32-mer and the 64-mer to the collapse of the core at high temperature, and its hydrophobic stability at low temperatures. The small chain showed an unrealistically small core size due to the few number of residues and folded through two transition, while the 64 monomer model exhibited a multistep folding process and showed a relative stability due to the large hydrophobic core.

In this section, we showed that for the same relatively long-ranged interaction potential, increasing the length of the chain from 64 to 128 and 256 residues results in changing the folding behavior from a two-step folding into a single-step folding transition. On possible explanation is that increasing the number of residues increases the effect of cooperativity in folding as the cooperativity in long chain is known to drive the collapse of the chain into the native state (Leonhard, Prausnitz, and Radke, 2003a; Kaya and Chan, 2000). Leonhard and coworkers (Leonhard, Prausnitz, and Radke, 2003a) argue that the 64-mer shows higher cooperativity and faster folding than the 27-mer due to the increase in the hydrophobic residues of the protein. The 27-mer exhibits very low cooperativity due to the number of
chain segments and lower surface-to-volume ratio compared to the 64-mer which are closer
to those of real proteins. Dobson and coworkers (Dobson, Sali, and Karplus, 1998) argued
that intermediates arise in the folding of lysozyme because of the lack of cooperative folding
in the entire protein from the collapsed state that is generated early in the folding process.
This can be explained by the cooperativity and the formation of a larger hydrophobic core
which scales with the number of residues in the model peptide. This is consistent with our
result as the 64-mer chain collapses at higher temperature than longer chains, and forms an
intermediate state during its folding to the low-temperature structure. In fact the 128-mer
and the 256-mer showed a larger core in the order of 2 and 3 respective to the core of the
64-mer.

In summary, the length of the chain plays a major role in proteins folding behavior. On
one hand, by increasing the number of residues, transitions merge into a single transition
from the extended state to the low-temperature state which is a behavior popular in pro-
tein chains. However, the computational intensity scales with the size of the chain making
the simulation of large chains in multichain system timely inefficient. Moreover, although
smaller chains appear computationally attractive, they lack the formation of a well-defined
core at low temperature and hence native conformations are highly unstable. Based on that,
a model chain of 64 monomers is chosen to use for further analysis. What is left is to opti-
mize the folding behavior of the 64-mer chain from a multiple step folding to a single-step
or a two-step folding process which are the most popular in proteins.
5.2 Effect of the overlap parameter ‘\(\sigma/l\)’

The next parameter under investigation is the overlap parameter defined as the ratio of the hard-core diameter to bond length (\(\sigma/l\)), which controls the chain’s stiffness. This parameter controls the stability of helices formed by homopolymer chains interacting through square well attractions (Magee, Vasquez, and Lue, 2006). The effects of stiffness are determined by examining models with varying \(\sigma/l\) while holding other model parameters constant. Simulations are run for \(\sigma/l\) equal to 1.0, 1.3, 1.6, and 1.9 for \(\lambda\) equal to either 1.5 or 1.1. The choice of \(\lambda = 1.1\) is not random, rather this specific range of interactions is optimized based on the study of its effect on the folding behavior of model chains presented in section 5.3. We investigated two relatively extreme values of \(\lambda\), corresponding to a short-range of interaction in which case simulations indicate a combined freezing/collapse transition. For the longer range of interaction (e.g \(\lambda = 1.5\)), the transitions are separated. Also, in the system with \(\lambda\) equal to 1.1, results of \(\sigma/l = 1.0\) are not included because the chain could not be equilibrated.

5.2.1 Model

In this section, we studied the effect of stiffness on the folding behavior of 64-mer chains. We first simulated four 64-mer chains in isolation with a range of interaction potential equal to 1.5 and stiffness equal to 1.9, 1.6, 1.3, and 1.0. Secondly, we simulated three single 64-mer chains with a range of interaction potential equal to 1.1 and stiffness equal to 1.9, 1.6, and 1.3. In both systems the sequence of the chain is the same as the one used in section 5.2.1 for the 64-mer chain with \(\lambda = 1.5\) and \(\sigma/l = 1.9\).

- **Sequence**: KEKSTAGRVASGVLDVSACVGLGIDTDLQGSPIAKLTFYGNKFNDVE ASQAHMIRWPNYTLPE

- **Other parameters**: 64-mer chains with an overlap parameter equal to 1.9, 1.6, 1.3, and 1.0 were simulated for \(\lambda\) equal to 1.5 and 1.1. The chain with \(\lambda\) equal to 1.1 and \(\sigma/l\) equal to 1.0 was not equilibrated and therefore not reported. By looking at the set of temperature used in replica-exchange, chains with lower \(\sigma/l\) values requires using more temperatures because the systems becomes harder to equilibrate.
• **Equilibration run:** For $\lambda$ equal to 1.5, $\sigma/l$ equal to 1.9, 1.6, and 1.3 required 30 million time-steps for equilibration, while chain with $\sigma/l$ equal to 1.0 required 45 million time-steps. For $\lambda$ equal to 1.1, chains with $\sigma/l$ equal to 1.9, 1.6, were equilibrated after 50, 65 and 75 million time-steps respectively to the lowest temperature.

• **Production run:** Independent runs of 20 million time-steps each were used for data analysis.

• **Temperature list:** In table C.1 found in appendix C, we present the set of optimized temperatures of each replica in the simulation calculated as shown in sections 4.2.1 and 4.2.2. Temperatures are input in reduced units as described in section 4.1.3.

### 5.2.2 64 monomer chain with $\lambda$ equal to 1.5

In this section we analyze the effect of the overlap parameter for 64-mer chains with $\lambda$ equal to 1.5. We extended the analysis of the previous 64-mer chain with $\sigma/l$ equal to 1.9 to three more 64-mer chains with $\lambda$ equal to 1.5 but for $\sigma/l$ equal to 1.0, 1.3 and 1.6.

Figures 5.30 and 5.32 are plots of the specific heat capacity as a function of dimensionless temperature for chains with $\sigma/l$ equal to 1.9, 1.6, 1.3 and 1.0. Figures 5.32 shows the change in $C_v$ between the chains with the highest and the lowest overlap parameter simulated. The number of transitions decreases with decreasing $\sigma/l$ values, as shown by the reduction in number of peaks in heat capacity plots. Chains with $\sigma/l$ equal to 1.0 and 1.9 exhibit two and four major transitions respectively, while chains with $\sigma/l$ equal to 1.3 and 1.6 exhibit 3 transitions. The types of transitions were identified for each chain based on contact maps, RMSD histograms, and RMSC histograms analysis. The folding behavior of the 4 chains with varying $\sigma/l$ is shown in figure 5.31.

The chain with $\sigma/l$ equal to 1.0 undergoes a high temperature disordered-disordered collapse transition followed by a disordered-ordered freezing transition. For chains with $\sigma/l$ equal to 1.3 and 1.6, the three peaks are identified as a disordered-disordered collapse transition followed by a disordered-ordered transition and a low-temperature ordered-ordered freezing transition. For $\sigma/l = 1.9$ there is a high temperature disordered-disordered collapse transition followed by a disordered-ordered transition forming the molten globule,
5.2. Effect of the overlap parameter ‘σ/l’

\[
\begin{align*}
C_v^* & = 1.9 \\
\frac{s}{l} & = 1.6 \\
\frac{s}{l} & = 1.3
\end{align*}
\]

Figure 5.30: A plot of specific heat capacity of the 64-mer as a function of temperature for chains with \( \lambda \) equal to 1.5 and \( \sigma/l \) equal 1.9, 1.6, and 1.3.

an ordered-ordered transition and a low-temperature ordered-ordered freezing transition forming the native structure shown in section 5.1.2.

It is clear that the overlap parameter has a tremendous effect on the folding behavior of peptide-models. Secondly, the peaks in the \( C_v \) shift to lower temperature with decreasing \( \sigma/l \). The collapse temperature of chain with \( \sigma/l \) equal to 1.9, 1.6, and 1.3 are at temperatures equal to 0.6, 0.5, and 0.4 respectively. The temperature of the collapse transition is correlated with the stability of the collapsed states. This provides an indication that decreasing the stiffness leads to a decrease in the stability of the collapsed state.

Figure 5.33 is a plot of the radius of gyration variation with dimensionless temperature for chains with \( \sigma/l \) equal to 1.9, 1.6, 1.3 and 1.0. For chains with \( \sigma/l \) equal to 1.6, 1.3, and 1.0 the \( R_g \) starts decreasing at the temperature of the high-T transition as expected for a collapse transition. For the stiffest chain, there is no data points between \( 0.7 < T^* < 2.2 \) since the \( R_g \) is not reweighted. The radius of gyration might be either constant or decreasing significantly across this temperature range which is above the collapse transition identified by the high-T peak in the \( C_v \) plot at \( T^* = 0.7 \). Secondly, the radius of gyration decreases with decreasing stiffness when compared at a given temperature. When reducing temperature from 2 to 0.7,
Figure 5.31: Diagram detailing the folding behavior of 64-mer chains with \( \sigma/l \) equal to 1.9 (a), 1.6 (b), 1.3 (c), and 1.0 (d). In each diagram, we show the main transitions and the properties of states across the transition with respective temperatures. N/D refer to not defined.
5.2. Effect of the overlap parameter ‘$\sigma/l$’

The size of the chain with $\sigma/l$ equal to 1.9 is decreased approximately by a factor of 2, while chains with smaller overlap parameter values retain a constant size. At low temperature all chains adopt a compact state with very close size ranging between 5 and 10.

5.2.3 64 monomer chain with $\lambda$ equal to 1.1

In this section we analyze the effect of the overlap parameter on the 64-mer chains with $\lambda$ equal to 1.1. The choice of $\lambda = 1.1$ is not random. The simulations investigating the effect of $\lambda$ on folding behavior in section 5.3 show that chains with $\lambda$ less or equal than 1.1 follow similar behavior (single or two-step folding), and because it is easier to equilibrate simulations with larger $\lambda$ values, we have chosen to study $\lambda$ equal to 1.1. If the effect was similar qualitatively to the one observed in chains with $\lambda$ equal to 1.5, then we show that it is independent of the other parameters, mainly the range of interaction potential.

Figure 5.34 contains plots of thermodynamic properties (specific heat capacity $C_v$, radius of gyration $R_g$, and internal energy $E$) as a function of reduced temperatures for three 64 monomer chains with $\lambda$ equal to 1.1 and $\sigma/l$ equal to 1.3, 1.6, and 1.9. All chains behave similarly: a single peak in $C_v$ at the temperature corresponding to the decrease in $R_g$ and
Chapters 5. Results and Discussion: effect of model parameters

Figure 5.33: Radius of gyration for the 64-mer with $\lambda$ equal to 1.5 and stiffness $\sigma/l$ equal to 1.9, 1.6, 1.3 and 1.0.

Reweighted internal energy. The nature of the transition and the states formed were identified for each chain based on contact maps, RMSD histograms, RMSC histograms, and SASA analysis. The results of this analysis showed that all chains are extended structures at high temperatures and fold into a native state with a hydrophobic core in a single-step transition with an all-or-none folding nature. The size and the energy of the chains increase by increasing stiffness at both high and low temperatures as shown in figures 5.34 (b) and 5.34 (c) respectively.

Figure 5.34 (b) shows the $R_g$ change of 3 chains as a function of temperature. At high temperature around $T^* = 0.5$, the size of the chains with $\sigma/l$ equal to 1.3 and 1.6 are similar while that of chain with $\sigma/l$ equal to 1.9 is almost two times larger. At low temperature, the three chains have a very similar size; the $R_g$ of the looser chain appears slightly lower.

Figure 5.34 (a) shows the change in the heat capacity with the overlap parameter. The peak in the $C_v$ of the stiffer chain is around $T^* = 0.225$ and extends at a temperature range of 0.21, while the $C_v$ peak of the chain with $\sigma/l = 1.6$ occurs at $T^* = 0.2$ covering an overall temperature range of 0.18 and the peak of the chain with $\sigma/l = 1.3$ is at a temperature approximately equal to 0.18 and extends to a temperature range of 0.16. The width of the
5.2. Effect of the overlap parameter ‘$\sigma/l$’

![Diagram of plots](image)

**Figure 5.34:** Temperature dependence of (a) Specific heat capacity, (b) Radius of gyration, and (d) Internal energy, for 64-mer chains with $\lambda = 1.1$ and $\sigma/l = 1.6$.

peaks are calculated at the minimum $C_v$ value at the lowest temperature. Accordingly as the overlap parameter decreases, the heat capacity peaks increase in height, become narrower, and shift to higher temperatures.
Figure 5.34 (c) describes the energy change as a function of reduced temperatures for three 64 monomer chains with $\lambda = 1.1$ and varying stiffness of 1.3, 1.6, and 1.9. At high temperature the energy of the loose chains ($\sigma/l$ equal to 1.6 and 1.3) is approximately equal to 0.1. That of chain with $\sigma/l = 1.9$ is slightly higher, about 1.7. The energy of the collapsed low-temperature state with $\sigma/l$ equal to 1.3 and 1.6 is -55 and that of chain with 1.9 is around -46. This indicates that the energy of the extended states does not vary much with stiffness, unlike that of the collapsed states. The previous analysis done in section 5.2.2 is still consistent but appears more clearly here as chains exhibit a single peak in $C_v$.

The temperatures of the peaks in $C_v$ correspond to the melting temperatures of the chains $T_m$ (or equivalently, the temperature where 50% of the structures are in the unfolded state yielding similar results (Doye, Sear, and Frenkel, 1998)). Accordingly, the change in the temperature of the transition with stiffness provides an indication that there is a direct correlation between the overlap parameter and stability of the collapsed states of the chain, and therefore the resistance to unfolding.

5.2.4 Short discussion

The overlap parameter influences the stability and the folding behavior of the model-chains in both systems. There are some similarities on the effect of the overlap parameters and some differences.

In both systems, there is a shift in the peak in heat capacities to lower temperatures by decreasing the stiffness. This indicates a decrease in the stability of the collapsed states. To explain the effect of stiffness on the collapse transition temperature, we look at the system with $\lambda = 1.1$ detailed in section 5.2.3, but the same analysis applies for the system with $\lambda = 1.5$. We consider the change in the Gibbs free energy difference between the extended and the collapsed state of a model-chain $i$ given by:

$$\Delta G_i = \Delta U_i - T_i \Delta S_i$$  \hspace{1cm} (5.1)

with $T_i$ indicating the collapse temperature of model-chain $i$ while $\Delta U_i$ and $\Delta S_i$ indicate the energy and entropy difference respectively between the collapsed and extended state of
model-chain $i$. At the collapse temperature, the free energy difference is equal to zero as there is an equal concentration of collapsed and extended states. The collapse temperature $T_i$ of the stiff chain 1 ($\sigma/l = 1.9$) and the loose chain 2 ($\sigma/l = 1.6$) is given by:

$$T_i = \Delta U_i / \Delta S_i \quad (5.2)$$

The energy difference between the extended and the collapsed state is defined as $\Delta U = U_c - U_e$, and $\Delta S = S_c - S_e$ respectively. By looking at figure 5.34 (a) and (c) the collapse temperatures $T_2 < T_1$, while the energy difference $\Delta U_1 = -46$, and $\Delta U_2 = -55$ implying that $\Delta S_2 < \Delta S_1$. Given that $\Delta S$ is negative, the previous equation implies that $\Delta S_2$ is more negative than $\Delta S_1$. The increase in the entropy difference by decreasing the stiffness decreases the entropic penalty for forming a collapsed state needed to produce a discontinuous expanded-to-compact transition with all-or-none characteristic. Given that $\Delta S_2 = S_{c,2} - S_{c,2}$, a more negative entropy difference for the looser chain can be caused by either a collapsed state with lower entropy and/or an extended state with a higher entropy.

In fact, the increase in the overlap parameter increases the stiffness of the chain by introducing restrictions on the allowed bond angles imposed by the excluded volume interaction between monomers separated by two bonds (Magee, Vasquez, and Lue, 2006). More interactions can be formed by a more flexible chain, whereas for a stiff chain, some interactions can not be formed as the chain can not fold back on itself. This reduces the number of possible configurations to be sampled at a given energy level which in turn reduces the entropy of the collapsed state. Moreover, a higher collapse temperature indicates an increase in stability of the collapsed state therefore stiffness correlates with the resistance to unfolding as shown by the shift in the melting temperature. This is valid for both systems.

The increase in the height of the heat capacity peak with decreasing overlap parameter indicates an increase in the latent heat of the transition with an increasing latent heat per monomer (Doye, Sear, and Frenkel, 1998). The energy difference between the extended and the collapsed state is defined as $\Delta U = U_c - U_e$, an increase in the energy difference would stabilize the loose polymer, but the entropy difference wins out as shown before.

Moreover, changing the overlap parameter changes the folding behavior of the model-chains in both systems: reducing the stiffness of the chain leads to a reduction in the number
of transitions in chains with $\lambda$ equal to 1.5 but only changes the shape of the peak in $C_v$ in chains with $\lambda$ equal to 1.1. The change in the folding behavior due to a change in stiffness can be related to the energetic constraints introduced with a change in stiffness. For chains with $\lambda$ equal to 1.5, reducing the stiffness of the chains leads to a decrease in the number of peaks as they grow taller and narrower, and shift to lower temperature. Since the peaks in $C_v$ are related to energy change and conformational transitions, reducing the stiffness possibly induces a discontinuity in folding (folding by a single or two-step transitions) simply because it costs less energy to fold or bend a loose chain on itself, and this can be done in a single transition.

On the other hand, the presence of stiff bonds reduces the entropy difference between the expended and collapsed states by restricting the amount of accessible conformations, this entropy cost can only be repaid if many monomers form contacts at the same time by compensating for the small contribution from the configurational entropy. The chain undergoes a conformational change without a specificity to the native contacts with possibly burial of hydrophobic core residues, forming any type of interaction to maximize the number of contacts and drive the collapse. These monomer-monomer contacts are not necessarily native interactions, and the chain needs other conformational transitions to rearrange themselves into more native-like structures. This continuity in folding is confirmed with the increase in the number of transitions with the overlap parameter and is consistent with many work done on homopolymers (Ruzicka, Quigley, and Allen, 2012; Taylor, Paul, and Binder, 2009a; Taylor, Paul, and Binder, 2009b) and simple HP models (Dobson, Sali, and Karplus, 1998).

In our system, the chain with $\sigma/l$ equal to 1.9 and $\lambda = 1.5$ exhibits a more continuous folding behavior compared to looser chains. The chain folds after four transitions as shown in the heat capacity plot in figure 5.32 while flexible chains with low overlap parameter fold with fewer transition. As the chain is cooled with time, there is a temperature point where the energy gain from forming contacts out-weights the entropy lost, and fewer transitions are required to fold the extended coil to the collapsed state, as shown in the $C_v$ plot of chains with $\sigma/l$ equal to 1.0 and 1.9 ($\lambda = 1.5$) in figure 5.32.

For the chains with $\lambda$ equal to 1.1 all chains exhibit a single transition so a reduction in the stiffness cannot possibly further reduce the number of transitions. The peaks in $C_v$ grow taller and narrower, occurring at a smaller temperature range by decreasing the stiffness
which can be explained in terms of energy and entropy change. The increase in the height of the peak with decreasing stiffness indicates an increase in the latent heat of transition as explained by Frenkel and coworkers (Doye, Sear, and Frenkel, 1998). This confirms the occurrence of a more discontinuous transition with the looser chains for the same reasons provided previously.

In short, the effect of the overlap parameter on the 64-mer chains with \( \lambda \) equal to 1.5 and 1.1 showed some common and some different features between the two chains possibly due to the effect of the range of interaction potential which is analyzed in the next section. Given that the stability of the chains increases with increasing stiffness, we chose an overlap parameter of \( \sigma/l \) equals to 1.9 for further studies.
5.3 Effect of range of interaction potential ‘λ’

In this section we study the effect of range of interaction potential on the folding behavior of model chains. The parameters optimized so far are the length of the chain with 64 residues and the overlap parameter $\sigma/l = 1.9$. Accordingly, we simulated 64-mer chains with $\sigma/l = 1.9$ and various ranges of interaction potentials equal to 1.03, 1.05, 1.07, 1.09, 1.1, 1.3 and 1.5. In order to show that the effect of range of interaction potential is independent of all parameters, we repeat the same analysis for the 64-mer chains with the same range of interaction potential but an overlap parameter equals to 1.3. Results of these simulations are shown in appendix A. We first present details specific to the models used in this section, next we present the main results, and we conclude with a short discussion.

5.3.0.1 Model

We simulate 7 chains with a range of $\lambda$ values. Simulation have been performed using replica exchange. Several runs of models with different starting configurations are performed in order to determine whether or not equilibration is reached. A detailed description of the model, interaction potential, and analysis tools common to all the simulations have been introduced in previous chapter 4.

Here we describe the key input parameters and specific simulation details of the seven 64-mer model related to this section.

- **Sequence**: KEKSTAGRVASGVLDsvACGVLGDIDTLQGSPIAKLKTfYGNKFNdVE ASQAHmIRWPNYTlPE
- **Other parameters**: The overlap parameter $\sigma/l$ is equal to 1.9 for the 7 chains with the range of interaction potential $\lambda$ equal to 1.03, 1.05, 1.07, 1.09, 1.1, 1.3, and 1.5.
- **Equilibration run**: All chains were equilibrated after a number of collisions within 20 million and 50 million time-steps; the number of time-steps to reach equilibration increased with decreasing $\lambda$.
- **Production run**: For each chain, independent runs of 20 million time-steps were used for data analysis.
5.3. **Effect of range of interaction potential \( \lambda \)**

5.3.1 **General overview**

Figure 5.35 contains plots of specific heat capacity \( C_v \) (a), radius of gyration \( R_g \) (b), and internal energy \( E \) (c) as a function of reduced temperatures for seven 64 monomer chains with various ranges of residue-residue interaction potential going from 1.03 to 1.5, all having an overlap parameter of \( \sigma/l = 1.9 \).

For \( \lambda=1.5 \), the \( C_v \) plot exhibits a dominant peak at \( T^* = 0.5 \) and a predominant peak at \( T^* = 0.13 \) while the peaks of the chain with \( \lambda = 1.3 \) are at temperatures around \( T^* = 0.34 \) and at \( T^* = 0.12 \). For chains with smaller range of interaction potential, the specific heat capacity displays a single well-defined peak at \( T^* = 0.2 \) for \( \lambda = 1.1 \), \( T^* = 0.19 \) for \( \lambda = 1.09 \), \( T^* = 0.18 \) for \( \lambda = 1.07 \), \( T^* = 0.165 \) for \( \lambda = 1.05 \), and \( T^* = 0.15 \) for \( \lambda = 1.03 \). The multitude of small peaks at low temperatures present in all chains are suggested to be a hallmark of glassy-like states that persist at low temperatures, and are similar to those found by Cellmer et al. (Cellmer et al., 2005b). There is a shift in the peaks to lower temperatures by decreasing the range of interaction from 1.5 to 1.03. As \( \lambda \) decreases, peaks in \( C_v \) grow taller and narrower. In fact, two distinctive behaviors are detected: chains with \( \lambda > 1.1 \) exhibit multiple transitions in their folding behavior while chains with \( \lambda \leq 1.1 \) exhibit a single transition with decreasing temperature. These two behaviors are consistent with the change in \( R_g \) of the chains and reweighted internal energy as shown in figure 5.35 (b) and (c) respectively. For temperatures between \( 0.25 < T < 2 \), the \( R_g \) and \( E \) of chains with \( \lambda \leq 1.1 \) are maximized and constant but reaches a near-minimum value for larger \( \lambda \) within the same temperature range. The size and the reweighted internal energy of the chains with \( \lambda \leq 1.1 \) starts decreasing at temperatures corresponding to the peaks in \( C_v \) and therefore the temperature of the collapse transition. The decrease is slightly more discontinuous with all-or-none characteristics as shown by an increase in the height and decrease in the width of peaks in \( C_v \), coupled with larger slopes in \( R_g \) and \( E \) plots. On the other hand, the size and the reweighted internal energy of the chains with longer range of interaction is continuously reduced from the high temperature slightly before the collapse transition occurs.

The structures along with the intermediates and the transitions have been identified in the folding behavior of the chains with different range of interaction potentials using contact maps, RMSD, RMSD\(_{CM}\), RMSC, and SASA analysis. These are detailed in the figure 5.36.
Figure 5.35: Temperature dependence of (a) specific heat capacity, (b) Radius of gyration, (d) and Internal energy of 64-monomer chains with $\sigma/l = 1.6$ and different $\lambda$. 
5.3. Effect of range of interaction potential ‘$\lambda$’

<table>
<thead>
<tr>
<th>Transitions</th>
<th>Temperature</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chains</td>
<td>Collapse</td>
<td>Melting</td>
</tr>
<tr>
<td>$\lambda = 1.5$</td>
<td>0.5</td>
<td>0.35</td>
</tr>
<tr>
<td>$\lambda = 1.3$</td>
<td>0.34</td>
<td>0.25</td>
</tr>
<tr>
<td>$\lambda = 1.1$</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>$\lambda = 1.09$</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>$\lambda = 1.07$</td>
<td>0.18</td>
<td>0.18</td>
</tr>
<tr>
<td>$\lambda = 1.05$</td>
<td>0.165</td>
<td>0.164</td>
</tr>
<tr>
<td>$\lambda = 1.03$</td>
<td>0.15</td>
<td>0.145</td>
</tr>
</tbody>
</table>

The chain with $\lambda$ equal to 1.5 exhibits the same folding behavior of the 64-mer chain studied in section 5.1.2 with four transitions: a high-temperature disordered-disordered collapse transition followed by a disordered-ordered transition forming a molten globule, an ordered-ordered transition and a freezing transition forming the native structure; this is shown in figure 5.36 (a). The chain with $\lambda$ equal to 1.3 exhibits three transitions: a high-temperature collapse transition forming a disorganized compact globule, a disordered-ordered transition forming a native-like structure with some fluctuations, and a freezing transition forming the native structure. The collapse and the freezing transitions are extensively studied in coarse-grained models for protein folding (Bratko and Blanch, 2001; Cellmer et al., 2005b; Cellmer et al., 2005c) and persist in the two chains with longer range of interaction potential.

On the other hand, the two intermediate transitions in the chain with $\lambda$ equal to 1.5 appear to merge into a single intermediate transition in the chain with $\lambda$ equal to 1.3. By further decreasing $\lambda$ from 1.3 to 1.03, all intermediate transitions disappear, the collapse and the freezing transitions come together into a single-step transition from the extended state to the native state, which appears as a sharp dominant peak in $C_v$. In figure 5.36 (c), we summarize the behavior of the chain with $\lambda = 1.1$, as smaller values of $\lambda$ follow the same trend.

The melting temperature is at a lower temperature than the collapse temperature for chains with $\lambda$ equal to 1.5 and 1.3 as shown in table 5.6. For $\lambda$ less or equal to 1.1, the melting and collapse temperatures are equal and they grow far apart as $\lambda$ is increased. This is another evidence that the number of transitions increases by increasing the range of interaction.
Figure 5.36: Diagram detailing the folding behavior of 64-mer chains with \( \lambda \) equal to 1.1 (a), 1.3 (b), and 1.5(c). In each diagram, we show the main transitions and the properties of states across the transition with respective temperatures.

Chapter 5. Results and Discussion: effect of model parameters
5.3.2 Short discussion

Changing the range of interaction potential affects not only the stability of the chains but also alters their folding behavior. Increasing $\lambda$ is coupled with a shift of the melting and the collapse temperature (if different) to higher temperatures indicating an increase in stability of the collapsed states as shown in table 5.6. Moreover, we found a critical value for the range of interaction potential above and below which the chains fold in different ways. This value is not very exact as the difference in $\lambda$ between chains is in the order of 0.2. For $\lambda \geq 1.3$, all chains fold through multiple transitions having the collapse and the freezing transitions common between them. By decreasing $\lambda$, intermediate transitions (transition other than the collapse and freezing transitions) merge first as shown in figures 5.36 (b) and (c) for chains with $\lambda = 1.3$ and $\lambda = 1.5$ respectively. Then the collapse and the freezing transitions come together and chains fold in a single step transition with all-or-none characteristics as shown by looking at chains with $\lambda = 1.1$ and $\lambda = 1.3$ in figure 5.36 (a) and (b) respectively.

Work on heteropolymers showed similar effect due to change in the range of interaction potential. Taylor and coworkers (Taylor, Paul, and Binder, 2009a; Taylor, Paul, and Binder, 2009b) found that for $\lambda > 1.05$ chains exhibit a two-step folding: a continuous collapse followed by a discontinuous freezing transition. For shorter range heteropolymer, the two transitions merge. The collapse transition is preempted by the freezing transition and there is a direct transition from an expanded coil to a compact crystalline state. The discrepancy in the quantitative analysis is due to the simpler model used by Taylor and coworkers, an off-lattice homopolymer; but the qualitative analysis rather show the same conclusion.

We try to generalize the effect of reducing the range of interaction potential on the folding behavior of the 64-mer independently of other parameters, including stiffness and length. Accordingly, we simulate chains with varying $\lambda$ with stiffness $\sigma$ equal to 1.3, in order to show that this effect is independent of $\sigma/l$. Results of these simulations are shown in appendix A and are in good agreement with previous findings.

In conclusion, chains with different ranges of interaction potentials exhibit different folding behavior. The effect of reducing the range of interaction potential is similar to that found by reducing the overlap parameters. In fact one possible explanation is that interactions between non-nearest neighbors can be established for larger $\lambda$, this should lead to a lower
energy of the collapsed disordered state and hence the collapsed state is more stable for higher $\lambda$ values. The loss in entropy is compensated by a continuous formation of contacts which leads to multiple transitions, mainly the collapse to a disordered state. On the other hand the cutoff in $\lambda$, for chains with $\lambda \leq 1.1$, probably corresponds to the range of the potential where interactions can only be formed between nearest neighbors. Chains therefore undergo a discontinuous transition at a relatively low temperature where all contacts are directly formed.

In summary, we explored in this chapter the effect of model parameters on folding behavior of model-chains. By fine-tuning the parameters ($\lambda$, $\sigma/l$, and number of beads $N$), many protein folding behaviors can be replicated. One of the folding behavior detected was the single-step transition from the extended state to the native state as shown for the 64-mer chains with $\sigma/l = 1.6$ and $\lambda \leq 1.1$. Secondly, a two-step folding appeared in the 32-mer with $\sigma/l = 1.6$ and $\lambda = 1.03$, and the 64-mer chain with $\sigma/l = 1.0$ and $\lambda = 1.5$. The third folding behavior includes the folding through three transitions as shown for the 64-mer chain with $\sigma/l = 1.3$ and 1.6, and $\lambda = 1.5$. Finally, we showed that the folding of the 64-mer chain with $\sigma/l = 1.9$ and $\lambda = 1.5$ occurs after four transition with the formation of the molten globule. These folding behavior have been detected in various proteins, but the folding with a single step transition or two transitions (collapse and freezing transitions) appear to be the most popular (Taylor, Paul, and Binder, 2009a; Ruzicka, Quigley, and Allen, 2012).

Accordingly, we look at the effect of parameters to choose an optimized model which exhibits the characteristics of real protein folding at low computational cost. First, reducing the number of residues in a chain induces folding through less number of transition and decreases the computational cost. On the other hand, the structures formed are unstable at low temperature which is not a characteristic of the native state in proteins. Therefore, the 64-mer chain appeared as a good optimization between computational cost and generic behavior of real proteins. This is consistent with the findings of Bratko et al. (Bratko and Blanch, 2001; Blanch et al., 2002; Bratko and Blanch, 2003). Moreover, increasing the overlap parameter and the range of interaction potential have consistent effect: both increase the stability of the chains at high and low temperature, and increase the number of transitions needed for chains folding. Similarly by reducing both parameters, chains appeared less stable at high and low temperature, and more difficult to equilibrate. This might be
related to the fact that increasing the range of interactions makes it more difficult to break the bonds, while increasing stiffness introduces conformational constraints to structures and therefore decrease the amount of accessible structures, making the search for the native state less computationally expensive.

Since we are looking for a model with certain stability but folds in either a single-step or a two-step folding, we choose the limit in the range of interaction below which chains fold in a single step, i.e. $\lambda = 1.1$, with large stiffness $\sigma/l = 1.9$ to maximize the stability of the native structure at low temperature, and its resistance to unfolding at high temperature. Accordingly, the model chosen is a 64 monomer chain with $\lambda = 1.1$, and $\sigma/l = 1.9$. This model-chain is the least computationally expensive, displaying all the characteristics found in real proteins.
Chapter 6

Results and Discussion:

low-temperature structures

For the upcoming analysis, we study a model with $N = 64$, $\sigma/l = 1.9$ and $\lambda = 1.1$ which was shown to exhibit generic protein-like behavior in chapter 5. Although we are primarily interested in the multichain system, we first characterize the folding behavior of five model-proteins in isolation with different sequences (chains 1, 2, 3, 4, and 5) that we use as a reference for studying these chains in multiple systems. We investigate the existence of a correlation between low-temperature conformational fluctuations and the stability of high-temperature collapsed states. We introduce definitions for these temperature-dependent fluctuations and put forward a new property, or order parameter, which discriminates between them. We also investigate the correlation between flexibility and stability of the chains at high and low temperatures.

6.1 Model

In this section, we simulate 5 off-lattice 64-mer model chains with residues interacting via the MJ potential. The five sequences are simply different amino-acid arrangement of the 64-mer used in the previous analysis, which is also the same sequence studies by Bratko et al. (Bratko and Blanch, 2001; Blanch et al., 2002; Bratko et al., 2006). In these sequences, the probability of an finding an amino acid in a protein sequence in nature is conserved. Simulation have been performed using replica exchange. For each sequence, several runs with different starting configurations are performed in order to determine whether or not
equilibration is reached. A detailed description of the model, interaction potential, and analysis tools common to all the simulations have been introduced in previous chapter 4. Here we describe the key input parameters and specific simulation details related to this section.

Details of the 64-mer model chain are listed below:

- **Sequence:**
  1. Sequence1: MIRWPNYTLPEKADSVACGVLGVLGLGKLTFASQAHDIDTLQGSP
  2. Sequence2: NDVEEKTAGLDSGVLGIDTLVACYTLPEQGSPIAKLKTGYGNK
  3. Sequence3: AFNYTLGDRVPEWKEKTAVLDSVLIDTLSPIAKLKTGYGNK
  4. Sequence4: WYTLFNSTGNEQASGGDSVGADRVPEKKTALSPIAKLKTGYNDV
  5. Sequence5: QGSPIAKLKTGFNKAACASQAHDIRPNKEKDAGYNVEYTLPE

- **Other parameters:** The range of interaction potential \( \lambda \) is set to 1.1, and the overlap parameter \( \sigma/l \) is equal to 1.9.

- **Equilibration run:** The system was equilibrated after 50 million time-steps to the lowest temperature.

- **Production run:** Two independent runs of 20 million time-steps each were used for data analysis.

- **Temperature list:** In table 6.1, we present the set of optimized temperatures of each replica in the simulation calculated as shown in sections 4.2.1 and 4.2.2. Temperatures are input in reduced units as described in section 4.1.3.

- **SASA analysis:** Using the SASA analysis, we define the residues with a low SASA (< 4) as buried residues. The choice of cutoff is subjective and depends on the progression of the average SASA value of the residues across the temperature range. Nonetheless,
6.2. From a single native structure to a native ensemble: the rational of the approach

Defining a native state for complex proteins is still a perplexing question. The old view suggested the existence of a stagnant three-dimensional ‘native’ structure at the lowest temperature. However, taking a static view of proteins and regarding them as stagnant chains at the lowest temperature it was clear that the residues have an average SASA either less than 4 or more than 10. The average SASA of the 5 chains at the lowest temperature \( T^* = 0.033 \) are shown in table 6.2. The residues in red are hydrophobic residues including glycine (G), alanine (A), valine (V), leucine (L), isoleucine (I), proline (P), phenylalanine (F), methionine (M), and tryptophan (W). Polar residues on the other hand are labeled in blue and include threonine (T), histidine (H), and tyrosine (Y).

### Table 6.2: SASA analysis of the 5 model-chains at \( T^* = 0.033 \) with a cut-off equal to 4. Polar residues are in blue while hydrophobic residues are in red.

<table>
<thead>
<tr>
<th>Chains</th>
<th>Sequence</th>
<th>Total core residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>chain 1</td>
<td>MIRWPNYTLP</td>
<td>36</td>
</tr>
<tr>
<td>chain 2</td>
<td>NDVEEKSTA</td>
<td>22</td>
</tr>
<tr>
<td>chain 3</td>
<td>AFNYTLCDR</td>
<td>39</td>
</tr>
<tr>
<td>chain 4</td>
<td>WYTLF</td>
<td>30</td>
</tr>
<tr>
<td>chain 5</td>
<td>QCSPIAK</td>
<td>35</td>
</tr>
</tbody>
</table>

### Table 6.1: Temperature list used for the 5 protein sequences in isolation.

<table>
<thead>
<tr>
<th>List number</th>
<th>Temperature (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.033</td>
</tr>
<tr>
<td>1</td>
<td>0.034</td>
</tr>
<tr>
<td>2</td>
<td>0.035</td>
</tr>
<tr>
<td>3</td>
<td>0.036</td>
</tr>
<tr>
<td>4</td>
<td>0.0367</td>
</tr>
<tr>
<td>5</td>
<td>0.0375</td>
</tr>
<tr>
<td>6</td>
<td>0.039</td>
</tr>
<tr>
<td>7</td>
<td>0.043</td>
</tr>
<tr>
<td>8</td>
<td>0.05</td>
</tr>
<tr>
<td>9</td>
<td>0.055</td>
</tr>
<tr>
<td>10</td>
<td>0.068</td>
</tr>
<tr>
<td>11</td>
<td>0.075</td>
</tr>
<tr>
<td>12</td>
<td>0.085</td>
</tr>
<tr>
<td>13</td>
<td>0.092</td>
</tr>
<tr>
<td>14</td>
<td>0.1</td>
</tr>
<tr>
<td>15</td>
<td>0.12</td>
</tr>
</tbody>
</table>

6.2. From a single native structure to a native ensemble: the rational of the approach

Defining a native state for complex proteins is still a perplexing question. The old view suggested the existence of a stagnant three-dimensional ‘native’ structure at the lowest temperature. However, taking a static view of proteins and regarding them as stagnant chains...
of amino acids is rather simplistic. Proteins undergo complex changes in space and time in order to keep cells functioning properly (Makowski et al., 2008). Some proteins, when in solution, exhibit conformational fluctuations (breathing) in their three-dimensional structures. Our simulations can possibly capture this breathing motion with the conformational fluctuations of the low-T states. In this section, we focus on the low-temperature states and the collapsed states behavior of model-chains 1, 2, 3, 4, and 5 to investigate the existence of a native ensemble. We report the folding behavior of the chains, which was studied using average contact maps, RMSD, RMSD$_{CM}$, and RMSC histogram plots to make sure it is consistent with the previous analysis. The change in radius of gyration, heat capacities and reweighted internal energy as a function of temperature was also investigated. All chains undergo a collapse from a high-temperature extended coil to a low-temperature compact structure. The low-energy states of sequences 1, 3, 4 and 5 form unique structures exhibiting a stable buried hydrophobic core, minor conformational fluctuations and contact map fluctuations. This shows that our model is good for studying protein aggregation. Results of these simulations are shown in appendix B. Accordingly, we investigate the possibility of defining the native structure as an ensemble of conformations rather than a single one.

Many reported various definitions of native conformation: Tang and Dill (Tang and Dill, 1998) suggest that a native protein is a Boltzmann ensemble of conformations with one conformation predominating a smaller populations of fluctuating conformations. They defined a conformational temperature $T_{cm}$, below which large fluctuations are frozen-out. All structures sampled below $T_{cm}$ form the native ensemble. This definition is in agreement with experimental observations on ribonuclease A. Another definition of native-ensemble is introduced by Cellmer et al. (Cellmer et al., 2005b) who calculated the chain’s free energy for several values of a structural overlap function at the folding transition temperature. The structural overlap function ($X$) is a measure of a snapshot’s similarity to the lowest energy structure. Then they calculated the free energy for several values of the structural overlap function at the folding transition temperature. The minimum at $X$ equal to 0.55 corresponds to compact non-native states. The local maximum at $X = 0.35$ is the thermodynamic barrier to folding. Thus, individual chains that have overcome this barrier, i.e., those with $X$ values less than 0.35, are in the native state. In another study, Cellmer et al. refer to a local energy maximum when calculating the free energy as a function of potential energy. Accordingly,
any chain that overcomes this energy barrier is a folded chain (Cellmer et al., 2005c).

In the traditional view, protein aggregation is caused by the association of the totally unfolded state(s) of proteins (Wang, 2005; Deyoung, Dill, and Fink, 1993; Stiger and Dill, 1993). However, there is overwhelming evidence that protein folding/unfolding intermediates are precursors in protein aggregation (Nagpal et al., 2015; Das et al., 2014; Fink, 1998). This more recent view is strongly supported by model predictions (Cellmer et al., 2005c; Bratko et al., 2007; Gupta, Hall, and Voegler, 1999), and by experimental data. Direct evidence for the transient association of partially folded intermediates during refolding has been obtained in small-angle X-ray scattering experiments on apomyoglobin (Eliezer et al., 1993; Eliezer et al., 1995), carbonic anhydrase and phosphoglycerate kinase (Semisotnov et al., 1996). Other experimental studies showed many partially-unfolded structures, such as scrapie amyloid (prion) protein (PrP27–30) (Safar et al., 1994), carbonic anhydrase B (CAB) (Wang, 1999), recombinant human growth hormone (rhGH) (Bam, Cleland, and Randolph, 1996), human lysozyme variants (Booth et al., 1997), P22 tailspike polypeptide (Schuler, Rachel, and Seckler, 1999), and phage P22 wild-type coat protein and its mutants (A108V, G232D, F353L) (Teschke, 1999).

Aggregation due to partial unfolding is explained by the exposure of a large part of the initially buried hydrophobic core in the unfolding intermediates (Wang, 2005; Wang, Nema, and Teagarden, 2010). The exposure of the highly hydrophobic core drives interchain interactions, and explains why aggregation occurs in bioprocessing stages. Shaking creates air/water interface; the hydrophobic property of air relative to water induces protein alignment at the interface, maximizing exposure of the hydrophobic residues to the air and initiating aggregation (Volin and Klibanov, 1989). Shearing also exposes the hydrophobic areas of proteins, causing aggregation. Many proteins easily aggregate during shaking or shearing, such as rFXIII (Kreilgaard et al., 1998b; Kendrick et al., 1998), hGH (Katakam, Bell, and Banga, 1995; Katakam and Banga, 1997; Bam et al., 1998), and insulin (Brange et al., 1997; Kwon et al., 2001).

The core not only plays a key role in the aggregation phenomena, but also induces low-temperature conformational fluctuations responsible of many protein catalytic activity. The protein can be therefore pictured as behaving somehow like a continuous elastic material.
that is referred to as the ‘breathing motion’. Chymotrypsin inhibitor 2 (CI2) is a well-doc-
dumented protein which displays the formation of a core and has been extensively studied both experimentally and using computer simulations. A disturbance in the core is seen to cause a change in the folding pathway. The folding patterns changes from three-state to two-state folding due to a single mutation in the buried residues (Khorasanizadeh, Peters, and Roder, 1996). In another study, Day and Daggett (Day and Daggett, 2003) found that while in the native state, the Chymotrypsin Inhibitor 2 protein exhibit a breathing motion characterized by a sliding of the helix and N-terminal strand relative to the sheet and active site loop. After undergoing this motion for some time, the protein reached the unfolding transition state, characterized by a weakened hydrophobic core and some loss of secondary structure. Once it passed through the transition state, the protein core rapidly became fully disrupted and the active site loop was highly distorted. A different study on BPTI, revisited by Kazmirski and Daggett in 1998 (Kazmirski and Daggett, 1998), showed that prior to partial unfolding of the chain, proteins lose specific core packing. Moreover, simulations indicate that protein G exhibits native conformational fluctuations at the termini of the fragment and in the turn itself, but the small hydrophobic core and inner hydrogen bonds are sufficient to maintain the structure (Roccatano et al., 1999). The work of Mccammon in 1984 (Mccammon, 1984) showed that structural fluctuations are essential for the biological activity of protein molecules. An example is ligand-binding molecules, like myoglobin or hemoglobin.

In summary, the core plays an essential role in three major mechanisms of proteins: the folding of the chain by the collapse of a hydrophobic core, low-temperature fluctuations, or breathing motion, and protein aggregation by exposure of the sticky core residues to other partially unfolded proteins. Accordingly, we introduce a new definition of a native-state ensemble based on our understanding of the core fluctuation which takes into account the native fluctuations at low temperatures. This property will allow us to distinguish between conformational fluctuations or breathing motion, and partial unfolding which induces aggregation. This ultimately divides the conformational space into two ensembles, folded and unfolded. Moreover, for small globular proteins, there is a strong correlation between aggregation propensity and the folding temperature implying the behavior of the protein at high temperatures correlates with the properties at room temperature. We explore whether or
not this correlation can be explained by the conformational fluctuations or breathing motion of the protein at room temperature.

6.3 Overview

6.3.1 Folding behavior

First we report the folding behavior of the 5 chains. We identify the main folding transitions and their types. We study the behavior of low-T states and try to explore the existence of an order parameter which distinguish between folded an unfolded structures. For small globular proteins, there is a strong correlation between aggregation propensity and the folding temperature implying the behavior of the protein at high temperatures correlates with the properties at room temperature. Typical experimental comparisons of stability of collapsed states are more often realized around the collapse temperature (Tang and Dill, 1998). Accordingly, we explore the relative stability of the collapsed states by identifying the temperature of the collapse transition.

We show plots of the temperature dependence of the specific heat capacity (6.1 (a)), radius of gyration(6.1 (b)), and reweighted internal energy (6.1 (c)). The folding behavior reported below was studied using average contact maps, RMSD, RMSD$_{CM}$, and RMSC histogram plots. The collapse transition was identified based on changes in $R_g$ and $E$.

The diagram shown in figure 6.2 summarizes the folding behavior of chain 1 by describing the change in all the relevant properties. The plot in heat capacity shows 2 peaks at temperatures equal to 0.24 and 0.19, and a shoulder peak at temperature equal to 0.16, indicating possibly that chain 1 folds through 2 or 3 transitions. The $R_g$ is an order parameter that is a good description of the collapse transition. Accordingly, the peak in heat capacity at $T^* = 0.24$ is identified as a disordered-disordered collapse transition. However the states formed below this transition have only 3 buried residues which is much less than expected. In fact, the burial of such small core does not compensate the entropy loss and cannot therefore drive the collapse of the chain into the low-T structure. The second peak in $C_v$ at temperature equal to 0.19 is another collapse transition (disorder-disorder transition) across which 22 core residues are buried. Usually these two transitions should merge in a
Chapter 6. Results and Discussion: low-temperature structures

Figure 6.1: Thermodynamic and structural properties of five 64-monomer chains with $\lambda = 1.1$ and $\sigma/l = 1.9$ as a function of reduced temperature. (a) Specific heat capacity, (b) Radius of gyration, (c) Internal energy.

Single disordered-disordered hydrophobic collapse transition. The third peak appears as a shoulder peak in heat capacity at temperature equal to 0.16 and was identified as a disorder-order freezing transition forming a well-defined structure with permanent contacts and a well-defined core with 36 residues.

Next we look at the folding behavior of chain 4 detailed in figure 6.3. The heat capacity
plot shows 3 peaks indicating 3 transitions. Using the $R_g$ and $E$ plots in figures 6.1 (b) and (c) respectively, we identify the collapse transition at temperature equal to 0.29 corresponding to the high-T peak in $C_v$. Similarly to chain 1, very few residues are buried across this transition, and the second peak in $C_v$ at $T^* = 0.23$ corresponds to a disordered-disordered hydrophobic collapse across which 28 core residues are formed. The last transition corresponds to the peak in $C_v$ at temperature equal to 0.15. This is a disordered-ordered freezing transition where the fluctuations in the disordered structure are frozen and contacts become more defined forming the native state.

The folding behavior of chain 5 is detailed in figure 6.4. The $C_v$ plot shows two peaks indicating two folding transitions. The high-T transition at $T^* = 0.22$ is identified as a collapse disordered-disordered transition while the low-T transition at $T^* = 0.22$ is identified as a disordered-ordered freezing transition. The folding behavior of chain 5 is closer to the expected behavior of the 64-mer model chains with $\lambda = 1.1$ and $\sigma/l = 1.9$ as the collapse from the extended state to the compact state should occur in a single transition not in two as shown for chains 1 and 4. In fact, for chain 5, twenty residues are buried across the collapse
transition, while in chains 1 and 4, three and five residues are buried respectively. One possible explanation is that the core formed across the collapse transition was large enough to drive the collapse of the structure in a single-step transition.

Figure 6.5 is a diagram of the folding behavior of chain 3. The heat capacity plot shows 2 main peaks at temperature equal to 0.14 and 0.20. The high-T transition exhibits a small shoulder peak at $T^* = 0.23$. Using the change in $R_g$ and $E$ in figure 6.1 the shoulder peak is identified as the collapse transition. This is a disordered-disordered transition forming a disordered globule with 7 buried residues. By decreasing the temperature to 0.20, the disordered globule arranges into a compact ordered structure with 31 buried residues constituting 80% of the core of the low-T structure. This ordered globule occurs at $T^* = 0.2$ which corresponds to the high-T peak in $C_v$. Below this temperature, the states are more ordered with larger hydrophobic core. At temperature equal to 0.144, the transition is identified as an ordered-ordered freezing transition to the native structure.

For chain 2, the behavior is very complex. In all the diagrams shown above, there is a red arrow indicating the melting/critical temperature where the average $<\text{RMSC}>$ is equal to
6.3. Overview

**Figure 6.4**: Sequence 5 folding behavior: diagram showing the thermodynamics and structural properties of transitions and intermediates of chain 5 at different temperatures on the heat capacity plot.

**Figure 6.5**: Sequence 3 folding behavior: diagram showing the thermodynamics and structural properties of transitions and intermediates of chain 3 at different temperatures on the heat capacity plot.
0.8, this is also revisited in the following section. For all the chains a collapse and a freezing transition were identified. In chains 1 and 4 the hydrophobic collapse extends across two peak because very few residues get buried across the first transition. For chain 3, the collapse transition appears as a peak with a shoulder in the heat capacity plot. On the other hand the collapse transition of chain 5 is identified with a single peak in $C_v$.

Moreover, all chains undergo a single disordered-ordered freezing transition, except for chain 3 where the transition is an ordered-ordered transition. One possible explanation is that the structure formed across the collapse transition has a relatively larger core than the other chains which makes it more resistant to higher temperatures. This idea is discussed in depth when we study the fluctuations in the following section.

6.3.2 Conformational fluctuations

In this section we investigate the behavior of the low-T states by examining the conformational and contacts fluctuations. We explore the relationship between low-T ensemble and high-T collapsed states by introducing a new order parameter related to core contact fluctuations.

Figure 6.6 is the average root mean square deviation in contacts of the core as a function of reduced temperature for 5 different chains. The core was identified using the SASA analysis, more details are found in section 6.1. Chains 1, 3, 4 and 5 have an $<\text{RMSC}>$ of 1 at the lowest temperature of $T^* = 0.033$ while the $<\text{RMSC}>$ of chain 2 is equal to 0.83.

The plot shows two different patterns, at low temperature the $<\text{RMSC}>$ slightly decreases with temperature, up to a temperature point where the $<\text{RMSC}>$ dramatically decreases towards zero. Although this temperature is specific for each chain, the drop occurs at a similar value of $<\text{RMSC}>$ equal to 0.8. This critical $<\text{RMSC}>$ value occurs at $T^* = 0.21$ for chain 3, $T^* = 0.176$ for chain 5, $T^* = 0.168$ for chain 1, $T^* = 0.15$ for chain 4, and $T^* = 0.055$ for chain 2. Chain 2 shows a slightly different behavior as the $<\text{RMSC}>$ of the lowest-temperature structure is already around 0.83. This is because the representative structure was chosen at a temperature ($T^* = 0.068$) where a binodal of ensemble of structures exists. However, the trend persists in chain 2, and the $<\text{RMSC}>$ is almost constant between the lowest temperature and $T^* = 0.055$, with an $<\text{RMSC}>$ equal to 0.81.
6.3. Overview

We next look at the correlation between the fluctuations in core contacts and contacts of the whole chain, i.e., between $<\text{RMSC}>$ and $<\text{RMSD}_{CM}>$ of the model-peptides. Figure 6.7 shows the average deviation in contacts of the chains ($<\text{RMSD}_{CM}>$) as a function of reduced temperature. At low temperatures, all chains exhibit few contact fluctuations and a constant minor decrease in contact deviation with temperature change. This can be attributed to the breathing motion of proteins. The $<\text{RMSD}_{CM}>$ of chains 3, 5, and 1 decreases from 1 to 0.7 at temperatures equal to 0.21, 0.176, and 0.168 respectively. Above these temperatures, the $<\text{RMSD}_{CM}>$ of the chains decreases rapidly towards zero. For chain 2 on the other hand, the $<\text{RMSD}_{CM}>$ has a value of 0.75 at the lowest temperature. This might be related to the $<\text{RMSC}>$ value of 0.83, which is lower than other chain as explained previously. The value is constant up to $T^* = 0.05$, where it decreases rapidly towards zero. Accordingly, all chains show critical value in average deviation in contacts above which the contacts start breaking rapidly with increasing temperature, the onset occurs at an $<\text{RMSD}_{CM}>$ equal to 0.7.

The critical value of $<\text{RMSC}>$ equal to 0.8 occurs at the same temperature of critical value of $<\text{RMSD}_{CM}>$ equal to 0.7. This suggests the possibility of choosing the deviation in contacts of the core as an order parameter to differentiate structures between folded and unfolded,
Chapter 6. Results and Discussion: low-temperature structures

Figure 6.7: Average root-mean-square deviation in all contacts \( \langle \text{RMSD}_{CM} \rangle \) as a function of temperature for chains 1, 2, 3, 4 and 5. That is any structure with an \( \langle \text{RMSD} \rangle > 0.8 \) or equivocally an \( \langle \text{RMSD}_{CM} \rangle > 0.7 \) can be considered folded. The temperature corresponding to the critical value of the order parameter is pointed out as a red arrow on the heat capacity plot for chains 1, 3, 4, and 5 in figures 6.2, 6.5, 6.3 and 6.4 respectively. The temperature corresponds to peaks (or shoulder peaks) in \( C_v \) indicating that the folded structures unfold across a transition which supports our claim.

For the chosen order parameter to be an adequate representation, we compare the temperature when the fraction of folded states is equal to 0.5 (according to this new definition of folded structure) to the heat capacity peak temperature. Figure 6.8 describes the percentage of folded structures as a function of reduced temperature, and shows of the melting temperature, \( T_m \), for all the chains. \( T_m \) is defined as the temperature at which the free energy of folding is zero, \( \Delta G(T_m) = 0 \), i.e., or equivalently the temperature at which 50% of the structures are in the folded state. The melting temperature of the 5 chains are summarized in table 6.3 along with the key temperature in the folding behavior of the chains.

Based on figure 6.8 and table 6.3, the melting temperature corresponds to the transition temperature found using the order parameter. Moving forward we will refer to this temperature...
6.3. Overview

Figure 6.8: Percentage of folded states of chains 1, 2, 3, 4 and 5 as a function of reduced temperature.

Table 6.3: Temperature of key transitions/properties of 5 optimized model-chains.

<table>
<thead>
<tr>
<th>Chains</th>
<th>80% &lt;RMSC&gt;</th>
<th>70% RMSD&lt;sub&gt;CM&lt;/sub&gt;</th>
<th>T&lt;sub&gt;freezing&lt;/sub&gt;</th>
<th>T&lt;sub&gt;melting&lt;/sub&gt;</th>
<th>T&lt;sub&gt;collapse&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chain 2</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.21</td>
</tr>
<tr>
<td>Chain 4</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.23</td>
</tr>
<tr>
<td>Chain 1</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.19</td>
</tr>
<tr>
<td>Chain 5</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0.18</td>
</tr>
<tr>
<td>Chain 3</td>
<td>0.21</td>
<td>0.21</td>
<td>0.14</td>
<td>0.22</td>
<td>0.23</td>
</tr>
</tbody>
</table>

as the critical temperature (no ambiguity with critical temperature in phase transitions). The red arrow in the diagrams showing the folding behavior of the chain corresponds therefore to the meting transition as well. The transition temperature calculated using the order parameter agrees with a peak in the heat capacity plot which corresponds to the melting temperature. Accordingly, the order parameter used to define a folded state, gives a transition at the peak in the heat capacity plot corresponding to the melting temperature. This proves that this order parameter is very efficient in defining whether or not a structure is in the folded or unfolded state.

The melting/critical temperature in chains 1, 4, and 5 corresponds to the disordered-ordered freezing transition. Accordingly, the freezing transition divides the structures between folded
and unfolded. For chain 3 the melting/critical temperature is equal to 0.21 which corresponds to the temperature of the peak in heat capacity higher than that of the freezing temperature as shown in figure 6.5. The freezing transition, which occurs at a lower temperature equal to 0.14, is an ordered-ordered transition. One possible explanation is that the increase stability of the native ensemble might be related to the relatively larger core with 39 residues constituting 61% of the chain. For structures at temperatures less than 0.14 the <RMSC> is constant at a value near 1, indicating no fluctuations in core contacts at all. The contact maps at temperature below 0.14 are formed of white and black squares only with no gray squares indicating that the structures either do not exhibit any kind of breathing motion or have very high stability. For temperatures between 0.144 (freezing-T) and 0.2 (critical-T) the <RMSC> and <RMSD_CM> decrease slightly but maintain values higher than 0.8 and 0.7 respectively indicating that structures still take part of the folded ensemble. At higher temperatures the fluctuations increase dramatically and the chain completely unfolds.

6.4 Short discussion

The first aim was to find an order parameter that allows us to distinguish between folded and unfolded states. This requires distinguishing fluctuations within the low-temperature structures or equivalently breathing motion (contact and conformational fluctuations) from partial unfolding of the collapsed states. We also investigate the possibility of comparing the stability of the low-T states with that of the collapsed states.

This is a stepping stone towards understanding the aggregation propensity of a chain from its low temperature behavior. At room temperature, proteins are biologically active (specific three-dimensional structure) in their native structure and exhibit a breathing motion. This breathing motion appeared as small conformational fluctuations of low-temperature structures due to a slight increase in temperature. On the other hand, the collapse temperature indicates the stability of the high-T collapsed state.

We use an order parameter to define whether or not a structure is in the folded or unfolded state, this order parameter definition is motivated by the plot of <RMSC> in core contacts. There is a critical value for the average fluctuations in core contacts, <RMSC> equal to 0.8, which occurs at a same temperature where the average deviation in contacts of the whole
6.4. Short discussion

chain $<\text{RMSD}_{CM}>$ is equal to 0.7. This temperature corresponds to the transitions in the folding behavior where 50% of the structures are folded. This also overlaps with the transition temperature according to the order parameter definition. This temperature also corresponds to transitions in the folding behavior, either freezing transition as seen in chains 1, 4, and 5, or collapse transition in chain 3.

Next we try to extrapolate a relation between stability of the low-T states and stability of the collapsed states. Although the key aspect correlating the core behavior to structural change is apparent in all model-proteins simulated, chains with different sequences exhibit different stability. In fact, the melting temperature and the collapse temperature are not the same for different model-chains. Chains with higher critical temperature have a more stable native ensemble, while chains with higher collapse temperature have a more stable collapsed states.

This is in agreement with an experimental study performed by Makowski et al. (Makowski et al., 2008) from Argonne National Laboratory and the Illinois Institute of Technology. They studied the breathing motions of a diverse group of five animal proteins (solutions of bovine hemoglobin, hen egg white lysozyme, hen egg white avidin, bovine serum albumen, and equine myoglobin) that represent a spectrum of size and structural differences. They found that the proteins varied widely in their kinetic behavior, and that each protein must be studied individually to learn its breathing habits under various cellular conditions, underscoring the danger of generalizing protein behavior from a limited sample.

By looking at table 6.3 we can extract trends in low-T states and collapsed states stability. Chain 2 is not a fair measure as it did not show a single structure at low temperatures. Accordingly, chain 4 has the lowest critical temperature (0.15) but the highest collapse temperature indicating that the most unstable low-T ensemble corresponds to a high stability of collapse states. For chains 1 and 5, the stability of the low-T states increase with increasing melting temperature to 0.16 and 0.17, while the stability of the collapsed states decreases by decreasing of the collapse-T to 0.19 and 0.18 respectively. On the other hand, although the native ensemble of chain 3 is the most stable, the collapse temperature is slightly higher than that of chain 5. This indicates that for three chains the stability of the low-T states scales inversely with the stability of the collapsed states. But this statement cannot be generalized as chain 3 behaves with a different trend.
In summary we defined an order parameter, the average root mean square in core contacts $<\text{RMSC}>$, which divides the structures across the folding behavior between folded and unfolded at a critical value of 0.8. This corresponds to a value in $<\text{RMSD}_{CM}>$ equal to 0.7. Also, this critical value occurs at the melting temperature of the chains for the models simulated, which in turn corresponds to the freezing transition in most models and to a higher temperature transition for unusually stable structures with exceptionally larger core. Also, we investigated the correlation between stability of the low-T and high-T states and found that some chains with more stable low-T ensemble correlates with unstable collapsed states.
Chapter 7

Results and Discussion: mutichain systems

In this section, we examine the aggregation behavior in terms of the transition temperatures known from previous chapter to distinguish folded structures from the unfolded ones. We identify which intermediates are aggregating and whether there are conformational changes upon aggregation. Then we map this conformational change to either conformational fluctuations within the native ensemble or partial/total unfolding above certain transitions. We also look at the correlation between chain stability and the aggregation propensity in multichain systems. For each sequence, a simulation box of 8 off-lattice chains in the native structure is allowed to run for more than $10^{10}$ collisions. Chains that form each multichain system are 8 replicated copies of the representative structure from the simulation of single chains in isolation i.e it is the structure with the most conformational similarities to all other structures at the lowest temperature (15000 structures in total). Snapshots are taken every $10^5$ collision and data are extracted for every snapshot. We visualize the evolution of chains in the system by relying on a set of order parameters: specifically the root mean square deviation in contacts of the core, interprotein interaction and intraprotein interactions are monitored. That way we obtain simultaneous measures for the resemblance to the native form, and for the degree of interprotein association.

In the first part, we will show results of one system (system 1 which is a multichain system of sequence 1) in order to correlate between different properties. In the second part we will correlate the behavior in each multichain system to the behavior of the corresponding chain in isolation (examine the aggregation behavior in terms of the transition temperatures,
which are clearly known from heat capacity plots) to test whether aggregation propensity can be correlated with behavior of chains in isolation, then we will compare the 5 multichain systems with each other.

7.1 Model

We simulate 5 multichain systems each with an off-lattice 64-mer model chains with residues interacting via the MJ potential. The system \( i \) corresponds to 8 copies of the \( i^{th} \) sequence simulated in chapter 6. The chain corresponds to the representative structure from the simulation of single chains in isolation, i.e., it is the structure with the most conformational similarities to all other structures (15000 structures in total). Snapshots are taken every \( 10^5 \) collision and data are extracted for every snapshot. Systems are not necessarily equilibrated as parallel tempering is not used. The results give insights into the initial kinetic steps involved in the association between the chains. A detailed description of the model, interaction potential, and analysis tools common to all the simulations have been introduced in chapter 4. The sequences are the same used in the previous chapter, for studying the low-temperature behavior. This allows extracting a relationship between the behavior of chains in isolation, and their propensity to aggregate.

Details of the 64-mer model chain are listed below:

- **Sequence:**
  1. Sequence1: MIRWPNYTLPEKADSVACGVLSGVLYGNKFNDVEEKSTAGRVA KLKTASQAHGIDTLQGSPI
  2. Sequence2: NDVEEKSTAGLDGSVGLDIDTLVACYTLPEQGSPIAKLKTGNYGNK FKRVASGVASQAHMIRWPN
  3. Sequence3: AFNYTLGDRVPEWKEFKSTAVLDVLDLSPAIKLTYGNKFNQV ESQAHMIRPGCGVAQASGG
  4. Sequence4: WYTLFNSGNEQASGGDSVADRVPEKKTALSPIAKLKTGNYGNKFSQAHMIRPGKAVLLIDCGV
7.2. A single multichain system

5. Sequence:

\begin{align*}
QGSPIAKLTYGKVFACASQAHMIRWPNKSTAGNDVEYTLP \\
ELDSGVLGDLKSVAGV
\end{align*}

- **Other parameters**: The range of interaction potential $\lambda$ is set to 1.1, and the overlap parameter $\sigma/l$ equal to 1.9.

- **Equilibration run**: No equilibration run. In this section, we are not looking at equilibrating our system, rather we are interested in the dynamics and mechanisms of aggregation. Therefore, data are collected as soon as the system starts interacting.

- **Production run**: Two independent runs of more than $10^{10}$ time-steps each were used for data analysis.

- **Temperature**: Systems are simulated at a single temperature equal to 0.033.

### 7.2 A single multichain system

To illustrate the relationship between core fluctuation and aggregation, in figure 7.1 we present time dependencies of root-mean-square deviation in contacts of the core (a), interprotein energy (b), and intraprotein energy (c).

At low temperatures all chains exhibit very low intramolecular energy and high intermolecular energy. The root-mean-square deviation in contacts of the core of all 8 chains in the system is equal to 1. This indicates that the chains are compact folded structures with a stable core which interact only through weak surface-surface interactions at low temperature. These are native state interactions which would occur between groups on the protein surface. Between $T^* = 0.033$ and $T^* = 0.14$ the intramolecular energy exhibits a large increase while the $<\text{RMSC}>$ and the interchain energy are almost constant. This indicates that intermediates across this temperature range are not prone to aggregation as individual chains undergo certain contact fluctuations in the core, but chains do not interact through parts with each other. As temperature is further increased, the intramolecular energy exhibits a large increase at $T^* = 0.14$ associated with a decrease in $<\text{RMSC}>$ from 0.9 to 0.57, and an increase in intermolecular energy indicating that unfolding leads to stronger intermolecular interactions.
Next we compare the behavior of the multichain systems to the behavior of chains in isolation for this specific sequence. The major decrease in \( \langle \text{RMSDC} \rangle \) occurs at 0.14 which corresponds to the freezing transition for the chain 1 in isolation which divides the structures between folded and unfolded. This is also the temperature at which the order parameter
7.3 Comparing multichain and isolated-chain systems

<RMSC> is equal to its critical value of 0.8 as shown by the diagram 6.2. Therefore the increase in fluctuations and reduction in intramolecular energy with increasing temperature below the freezing temperature which is not affected by intermolecular association or aggregation. On the other hand, structures above the freezing temperature are unfolding quickly with an increase in interchain energy. This indicates that unfolded structures are prone to aggregation as association only occurs significantly once the unfolded states are formed.

In short, the contact fluctuations in the core corresponding to the breathing motion of the isolated chain do not cause the formation of interchain contacts. This is another evidence that the order parameter provides a good indication of the native ensemble as aggregation does not occur from the native states. This is in good agreement with the data in multichain system. As the deviation in the core contacts increases and the <RMSC> drops below 0.9-0.8, the intermolecular energy starts increasing significantly, indicating the formation of interchain contacts, or aggregation. This from one hand supports the existence of a correlation between the core compactness, and aggregation propensity, and on the other hand it shows that the ensemble of structures above the freezing transition promotes aggregation. One possible explanation is that the unfolding of the core causes exposure of hydrophobic residues thereby promoting the formation of interchain contacts and induces aggregation.

7.3 Comparing multichain and isolated-chain systems

The behavior of multichain system 1 was described in the previous section and correlated with the behavior of chain 1 in isolation. The same analysis applies for systems 2, 3, 4, and 5, this is detailed in appendix D where we show the temperature dependence of intermolecular energy and root mean-square deviation in core contacts for the 5 systems. The starting configuration corresponds to the representative structure obtained from clustering the structures at a temperature of 0.033. Here we look at these properties averaged over the 8 replicas of a given sequence in a system.

In figure 7.2, we show 2 plots of the temperature dependency of <RMSC> and <E_{inter}> averaged over the 8 chains per system. Thus for every system we report one <RMSC> data set and one <E_{inter}> data set with an error bar showing the standard deviation in the property corresponding to the 8 chains. We compare the change in <RMSC> and <E_{inter}> with
temperature between the 5 systems, and for every system we can explore the correlation between the $<\text{RMSC}>$ and $<\text{E}_{\text{inter}}>$ as the temperature axis overlaps as shown in the figure. Moreover, in order to identify the nature of intermediates aggregation we refer to the melting temperatures $T_m$ above which chains are considered unfolded.

All the chains exhibit two behavior, and there is a temperature below which fluctuations in $<\text{RMSC}>$ and $<\text{E}_{\text{inter}}>$ do not induce aggregation. This temperature is very close to the transition temperature in single chains where the order parameter is equal to its critical value. In
7.3. Comparing multichain and isolated-chain systems

Fact at low temperature, the gradual decrease in $<\text{RMSC}>$ correlates with minor fluctuations in $<\text{E}_{\text{inter}}>$. The interactions occur possibly through some parts of the chains. However, at temperature above these specific transitions, the $<\text{RMSC}>$ drops significantly, and the $<\text{E}_{\text{inter}}>$. decreases towards a minimum indicating that structures with large conformational changes are associated with aggregation.

The correlation between the melting temperature transition and the aggregation propensity appears the clearest with systems 1 and 3. For system 3, the $<\text{E}_{\text{inter}}>_{T^*=0.21}$ is the largest indicating minimum interchain contacts. At temperatures above 0.21, the average interchain energy decreases rapidly which corresponds to the temperature of the unfolding transition. Similarly for system 1, the large decrease in $<\text{E}_{\text{inter}}>_{T^*=0.16}$ is associated with a decrease in $<\text{RMSC}>$ at temperatures above the unfolding transition (and critical order parameter temperature). This shows that aggregation is correlated with intermediates exhibiting conformational changes at temperatures above the unfolding transitions associated with $T_m$.

Moreover, chain 4 was shown to have the least stable native ensemble and the most stable collapsed states. Here, if we disregard chain 3 due to its uncommon high stability associated with the largest core, chain 4 appears to be the most resistant to aggregation with the lowest $<\text{E}_{\text{inter}}>_{T^*=0.18,0.21}$.

In figure 7.3 we present five plots, each with two sets of data. In one set we show the $<\text{RMSC}>$ averaged over 8 chains in the multiple-chain system at a given temperature with a standard error on the average, in the second set we present the $<\text{RMSC}>$ of single chains in isolation corresponding to the multiple chain systems. The change in $<\text{RMSC}>$ with temperature of the 8 model peptides per system before averaging in the multiple system are shown in figure D.2 in appendix D.

The general behavior in $<\text{RMSC}>$ change with temperature appears to be similar in both single and multiple-chain systems. The $<\text{RMSC}>$ change in multichain systems shows two behaviors, there is a temperature corresponding to a large decrease in $<\text{RMSC}>$ towards a minimum similar to the behavior of chains in isolation. This temperature was previously correlated with the melting transition of single chains at temperature where the $<\text{RMSC}>$ is equal to 0.8.
Chapter 7. Results and Discussion: multichain systems

At low temperatures (within the native ensemble) the \(<\text{RMSC}>\) of chain in multiple peptides appears to be slightly higher than those of chains in isolation. One possible explanation is that structures in the low-T states are stuck in certain local energy minima because the multichain systems are not equilibrated as we are not using parallel tempering.

The melting temperature, where the fluctuations in \(<\text{RMSC}>\) becomes large, is slightly smaller in multichain systems than in chains in isolation. At temperatures above the melting, the decrease in \(<\text{RMSC}>\) in mutichain system is larger than that of chains in isolation.
This can possibly be justified by the fact that when the concentration of the peptides is increased, fluctuations within the collapsed states promote more unfolding or aggregation.

In systems 1 and 5, the <RMSC> for the multiple chain and chains in isolation appear to overlap across the temperature range. However, in system 1 the contacts in structures below the melting transition appear to be less fluctuating.

In systems 2 and 4, the two plots of <RMSC> in single and multiple system follow the same trend, but the degree of contact fluctuations appears to be smaller in the multichain system as the <RMSC> is always slightly larger.

In system 3, at temperatures below the melting transition (0.14) the <RMSC> of the chains in isolation and in multiple system are almost equal. At higher temperatures the peptides <RMSC> in multichain system is much lower than that of chains in isolation.

7.4 Short discussion

There is a clear correlation between the folding behavior of chains in isolation, and the unfolding/aggregation of the chains in multiple-chain systems. We showed that it is possible to divide the structures forming by the folding of the chain into the folded or native ensemble and the unfolded ensemble based on an order parameter, the <RMSC>. Moreover, for 4 chains out of 5, this temperature matched the melting disorder-ordered transition while for the last model-chain (chain 3) it correspond to a temperature slightly lower than the melting temperature because of the larger core of the chain. Moreover, the stability of the low-T states was examined based on the temperature corresponding to the critical value of the order parameter. The stability of the collapsed states was examined based on the collapse temperature and it was found that some chains with unstable native ensemble have a higher collapse temperature.

In the multichain system, the data were as expected. First, the behavior of the chains is similar to that of chains in isolation. Based on the data of <RMSC> and <E_{inter}>, conformational fluctuations at temperatures below the unfolding/melting temperature (corresponding to
the transition where the order parameter is equal to its critical value) do not induce aggregation and are correlated with breathing motion of the chains. At higher temperature unfolding induces interchain contacts irreversibly. Therefore partially unfolded intermediates, which take part of the unfolded ensemble, are the main precursor of aggregation.
Chapter 8

Results and Discussion: square-well fluids

In this chapter, we investigate the supercritical behavior of square-well fluids with different ranges of interaction potential using molecular dynamics (MD) computational simulations for systems up to 10976 particles. We developed a more efficient way of calculating the critical temperature of these square-well fluids, specifically for strongly short-ranged systems which are especially difficult to simulate as they exhibit glassy behavior. Other methods require using simulations at lower temperature in the two phase region and hence will be more difficult to equilibrate. The critical temperatures of fluids with $\lambda$ equal to 1.005, 1.5 and 2 are found to agree with literature. Identifying the critical temperature and how it depends on the pair potential is very important in bioprocessing and bioformulation as opalescent solutions that occur near to the critical temperature need to be avoided (Mason et al., 2011).

Here we focus on the adhesive-sphere model, or what is collectively known as the sticky-sphere model, for fluids with $\lambda = 1.005$. The range of interaction in a protein solution is very small compared to the size of proteins, and can be considered to occur on the surface of the proteins. Thus for very small range of interaction, a SW fluid consisting of $N$ spheres simulates a solution with $N$ proteins. The empirical method derived could then be applied to more realistic short-ranged potentials for proteins which could be patchy spheres or possibly non-spherical shapes interacting through anisotropic interactions.

First we give an overview about percolation transitions and their relevance to the new method proposed. The isotherms of square-well fluids with a ranges of interaction equal to 1.005, 1.5 and 2 are calculated as a function of density. Molecular dimers, trimers . . . etc.
... up to large clusters, or droplets, of many adhesive-sphere atoms are unambiguously defined. We finally introduce a new method to calculate the critical parameters (temperature and pressure) by defining percolation transitions at the molecular level that bound a supercritical region and compare the critical values to literature.

8.1 Background

The thermodynamic property rigidity, defined as \((dp/d\rho)_T\), where \(p\) is pressure, \(\rho\) is density and \(T\) is temperature, is the work required to reversibly increase the density of a fluid. This state function decreases with density for a gas and increases with density for a liquid. Thereby a dense supercritical region with liquid like properties, and dilute supercritical region with gas-like properties can be distinguished from each other by their rigidity. Rigidity shows a remarkable symmetry between liquid and gas phases due to an equivalence in number density fluctuations, arising from available holes in the liquid state, and molecular clusters in the gas.

Along supercritical isotherms, the densities at which molecular clusters of occupied sites in the gas phase, or clusters of unoccupied voids in the liquid phase, first become macroscopic, are referred to as percolation transitions. Percolation transitions in model fluids of hard-spheres and square-well molecules are well defined by a characteristic distance, which unambiguously divides configuration space volume into ‘sites’ and ‘holes’. An illustration of the available-volume and excluded-volume is shown in figure 8.1.

![Figure 8.1: Two-dimensional illustration of available volume (Va: yellow) and excluded volume (Ve: red plus black).](image)

The percolation transition occurs at densities where small clusters of occupied sites in the gas phase, or unoccupied holes in the liquid phase, first span the whole system. These are
denoted by PB and PA respectively, referring to ‘bonded cluster’ (sites) and ‘available volume’ (holes) respectively. An illustration of percolation transitions is shown in figure 8.2. The percolation loci (PL) can, nonetheless, be defined thermodynamically along supercritical isotherms by the conditions (Woodcock, 2013):

\[
\begin{align*}
PL \left\{ \\
\rho < \rho_{PB} \quad & (d^2p/d\rho^2)_T < 0 \quad \text{gas} \\
\rho_{PA} < \rho < \rho_{PB} \quad & (d^2p/d\rho^2)_T = 0 \quad \text{mesophase} \\
\rho > \rho_{PA} \quad & (d^2p/d\rho^2)_T < 0 \quad \text{liquid}
\end{align*}
\] (8.1)

here the mesophase corresponds to the supercritical region between the percolation transition at constant rigidity. Rigidity \([(dp/d\rho)_T]\) is inversely proportional to fluctuations in the atomic number density. For a gas of attractive atoms, rigidity decreases with density because fluctuations increase with polymerisation, whereas for a liquid, rigidity increases with density because the voids become fewer as density increases. This condition holds true for the subcritical isotherms of liquid and gas when \(T < T_c\).

Figure 8.3 is a plot of the pressure versus density for experimental points of supercritical argon, obtained from measurements of Gilgen et al. (Gilgen, Kleinrahm, and Wagner, 1994) and presented by Woodcock (Woodcock, 2013). The extended straight lines are fitted to the
points in the intermediate region, referred to as the mesophase in equation 8.1, dividing the supercritical isotherms into three different parts. Actually this mesophase is bounded by the available-volume and the bounded-cluster percolation transitions, respectively PA and PB. Inside this region, the experimental points are fitted to a perfectly straight line. These points have been reported to a six-figure accuracy. Thus the percolation transitions bound the existence of dense supercritical region with liquid like properties, and dilute supercritical region with gas-like properties.

![Figure 8.3: Experimental data points for the $p$-$\rho$ isotherms on the supercritical argon from the results of Gilgen and coworker (Gilgen, Kleinrahm, and Wagner, 1994).](image)

The main objective of the work is to introduce a new method for calculating critical parameters. We investigate the location of the mesophase as a function of range of potential and we show that the critical temperatures can be determined by extrapolating the slopes in the mesophase to 0. This method does not require the simulation of any phase behavior below the critical temperature and as such could provide a quicker method for obtaining critical temperatures. It also avoids more computationally intensive phase behavior simulations which will be more difficult to equilibrate as they occur at lower temperatures.
8.2 Model

All the results are shown for DMD simulation of square-well systems of \( N = 10976 \) particles, (14x14x14x4) fixed initially on a FCC lattice. We simulate supercritical isotherms in three model-fluids with different range of interaction potential equal to 1.005, 1.5, and 2.

- \( \lambda = 1.005 \): this is referred to as the sticky-sphere limit. In theory at the sticky sphere limit, \( \lambda \to 1 \), but it is well-known that models with \( \lambda = 1.005 \) capture the essential behavior observed in the sticky-sphere limit. In this limit a particle might simulate a protein due to the very large particle size to range of interaction ratio.

- \( \lambda = 1.5 \): this corresponds to the a characteristic range of interaction describing molecules in a vacuum and therefore can be referred to as the real-molecules limit, as schematized in figure 8.4.

- \( \lambda = 2 \): for hard spheres, the exclusion volume of a sphere of hard core diameter \( 1\sigma \) is \( 2\sigma \). And therefore \( \lambda = 2 \) is the excluded-volume limit for hard sphere and is a special case for square-well simulations.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{fig8_4.png}
\caption{Schematization of the square-well \( \lambda = 1.5 \) range of interaction, as the best approximation for interactions in real molecules.}
\end{figure}

Each isotherm is simulated with more than 112 state point for densities ranging between 0 and 1. All systems are equilibrated. A system is considered equilibrated if two successive runs for a certain amount of collision yield the same isotherms.

The square-well hamiltonian is the simplest model of Var der Waals molecular fluid and has been extensively investigated as a test of the augmented Var der Waals equation.
If $i$ and $j$ are monomers separated by a distance $r$, they will interact via square-well potential:

$$u_{i,j}(r) = \begin{cases} 
\infty & \text{for } r < \sigma \\
-\epsilon & \text{for } \sigma < r < \lambda \sigma \\
0 & \text{for } \lambda \sigma < r 
\end{cases}$$ (8.2)

where $\sigma$ is the hard-sphere diameter, $\epsilon$ is the square-well depth, and $\lambda \sigma$ is the range. The reduced thermodynamic states are then temperature $T^* = k_B T / \epsilon$ and density $\rho^* = N^3 / V$, where $k_B$ is Boltzmann’s constant and $N/V$ is the number density.

### 8.3 Overview

#### 8.3.1 Model fluids

Figure 8.5 shows pressure-density isotherms of 3 model fluids. In figure 8.5 (b) 11 isotherms are obtained for a range of interaction of $\lambda = 1.5$, with 1232 state point in total along 5 near-critical isotherms between 1.2 and 3. The system is equilibrated after $6 \times 10^8$ collisions, and every point is an average of $10^8$ collisions, after equilibrating the system. Figure 8.5 (c) shows pressure-density isotherms for 19 isotherms listed in the figure legend. The isotherms are obtained for a range of interaction of $\lambda = 2$, with 2128 state point in total along 7 near-critical isotherms between 2.45 and 3. The system is equilibrated after $6 \times 10^8$ collisions, and every point is an average of $10^8$ collisions. The two percolation transitions can be observed on all the supercritical isotherms as a discontinuity in the slope of the $p(\rho)$ isotherms. All the isotherms show three distinctly different regions of behavior for the state function:

- In the low density region (gas phase), for $\rho < \rho_{PB}$, the compressibility $(dp/d\rho)_T$ decreases with density.

- In the mesophase $\rho_{PB} < \rho < \rho_{PA}$, the compressibility $(dp/d\rho)_T$ is constant and is shown by the straight line. In all isotherms this is mainly the region with the highest number of state points, where they appear very condensed.
• In the high density region (liquid phase), for $\rho > \rho_{PA}$, the compressibility $(dp/d\rho)_T$ increases with density.

Figure 8.5: Mean pressures from simulations of the square-well fluid with $\lambda = 1.005$ (a), $\lambda = 1.5$ (b), and $\lambda = 2$ (c).

Figure 8.6 shows two expanded plots of figure 8.5 obtained by zooming at the region $0.4 < \rho < 0.5$ for the six first isotherms for model fluid with $\lambda = 1.5$ (a), and four near-critical (super-
and sub-critical) isotherms at $T^* = 2.45, 2.5, 2.6, 2.7$ and $2.8$ for model fluid with $\lambda = 2$ (b).

The figures show that there is a region for densities between 0.39 and 0.48 where the slope is linear in both systems, also the critical isotherm in the mesophase is a straight line with a zero-slope. The data points are fitted with straight lines which reflects this linearity within a certain density region. The sticky-sphere model will be discussed in more details in the upcoming analysis.

![Graph](image-url)

**Figure 8.6**: Expanded plot of (a) 5 supercritical isotherms shown in figure 8.5 (b) for a region within the mesophase of model-fluid with $\lambda = 1.5$, and (b) 17 supercritical isotherms shown in figure 8.5 (c) for a region within the mesophase. The data points are fitted with straight lines which reflects a linearity within this density region. The dotted line is fitted to the critical isotherm.
8.3. Overview

8.3.2 Sticky-sphere model

8.3.2.1 Determination of critical parameters

Next we present detailed analysis of the SW fluid with $\lambda = 1.005$, and introduce a new method for calculating critical parameters. The pressures along 7 supercritical isotherms are shown in figure 8.7. Every state point is equilibrated independently for 10 million collisions and then simulated for 1 billion collisions with ensemble averaging of thermodynamic properties. Each isotherm shown in figure 8.5 (a) is simulated with more than 70 state point for densities between 0 and 1, with a total of 700 state points for 8 isotherms going from $T/T_c = 1$ until $T/T_c = 1.1$. Calculation of the critical temperature will follow next. As temperature increases, the range of constant rigidity gradually decreases and appears to be the smallest at $T^* = 0.220$. This range disappears at higher temperature simulated, specifically it extends to $T^* = 0.225$ which corresponds to $T/T_c = 1.12$.

![Figure 8.7: Supercritical isotherms of $\lambda = 1.005$ adhesive sphere SW model; the critical isotherm is $T_c = 0.2007$ showing pressure as a function of density from $T_c$ to $1.1T_c$. The data points are fitted to a straight line shown as a dotted black line on the figure indicating the linear area and denoting the densities of the percolation transitions extracted from the constant rigidity $(dp/d\rho)_T$ shown in figure 8.8.](image)

The linear combination rule can be used, along with known experimental $\rho(p, T)$ data in the near-critical mesophase region, to obtain critical parameters without resorting to simulation
Table 8.1: Properties of supercritical isotherms of an adhesive-sphere square-well fluid \((\lambda = 1.005)\).

<table>
<thead>
<tr>
<th>(T^*)</th>
<th>Slope (R_m)</th>
<th>Intercept (p_0)</th>
<th>PB density</th>
<th>PB pressure</th>
<th>PA density</th>
<th>PA pressure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0.4460</td>
<td>-0.0471</td>
<td>0.35</td>
<td>0.109501382</td>
<td>0.42</td>
<td>0.140864156</td>
</tr>
<tr>
<td>0.22</td>
<td>0.1725</td>
<td>0.0033</td>
<td>0.32</td>
<td>0.058580453</td>
<td>0.40</td>
<td>0.072286274</td>
</tr>
<tr>
<td>0.217</td>
<td>0.1479</td>
<td>0.0068</td>
<td>0.32</td>
<td>0.054145429</td>
<td>0.42</td>
<td>0.068959722</td>
</tr>
<tr>
<td>0.215</td>
<td>0.1294</td>
<td>0.0097</td>
<td>0.3</td>
<td>0.048435921</td>
<td>0.41</td>
<td>0.062755033</td>
</tr>
<tr>
<td>0.212</td>
<td>0.1032</td>
<td>0.0135</td>
<td>0.32</td>
<td>0.046451419</td>
<td>0.43</td>
<td>0.057718793</td>
</tr>
<tr>
<td>0.21</td>
<td>0.0786</td>
<td>0.0184</td>
<td>0.32</td>
<td>0.043268979</td>
<td>0.46</td>
<td>0.054366218</td>
</tr>
<tr>
<td>0.206</td>
<td>0.0461</td>
<td>0.0203</td>
<td>0.28</td>
<td>0.03305980</td>
<td>0.51</td>
<td>0.043856251</td>
</tr>
<tr>
<td>0.2007</td>
<td>0.0009</td>
<td>0.0276</td>
<td>0.26</td>
<td>0.02780010</td>
<td>0.47</td>
<td>0.02798837</td>
</tr>
</tbody>
</table>

data in the two-phase region. Every isotherm in figure 8.5 (a) or figure 8.7 obeys the linear equation for the pressure

\[
p_m = p_0 + R_m \rho
\]

where \(R_m\) is the constant isothermal rigidity \((dp/d\rho)_T\) within the mesophase. It is independent of density, and as shown in figure 8.8, can be easily parameterized. Slopes and intercepts of the seven supercritical isotherms in figure 8.8 obtained directly from the molecular dynamics simulations are collected in table 8.1. These parameters can be used to estimate directly a critical temperature \((T_c)\) and critical pressure \((p_c)\), respectively, and also to obtain the pressures and densities, of the percolation loci PB and PA.

For densities both below PB, and densities above PA, on every isotherm, the departures in pressure \(\Delta p = (p - p_m)_{T, \rho}\) are found to be quadratic in density. A plot of density against the modulus \(|(p - p_m)|^{1/2}\) gives a straight line that interpolates at \(\Delta p = 0\) to obtain the percolation densities at PA and PB and, hence also the pressures using equation 8.3, all as given in table 8.1.

The slopes of the mesophase rigidities decrease linearly with temperature in the vicinity of \(T_c\), which is then obtainable from the measurements of the rigidity constants in the mesoscale region by extrapolation of the slopes of the isotherms against temperature to zero as shown in figure 8.9. \(T_c\) is found to be 0.2007 ± 0.002.

Likewise, a critical pressure can also be obtained from the supercritical isotherms by plotting the intercept of the linear plots \((p_o\) in equation 8.3) against \(T - T_c\) and interpolating to
Figure 8.8: Supercritical isotherms of $\lambda = 1.005$ adhesive sphere SW model; the critical isotherm is $T_c = 0.2007$ showing pressure as a function of density from $T_c$ to $1.1T_c$.

Figure 8.9: Determination of $T_c$: plot of slopes in figure 8.8 against $T$ to determine $T_c$. 
Table 8.2: Critical temperature calculated using the new empirical method for SW fluids with \( \lambda \) equal to 1.005, 1.5, and 2.

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>( T_c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.005</td>
<td>0.2007</td>
</tr>
<tr>
<td>1.5</td>
<td>1.1801</td>
</tr>
<tr>
<td>2</td>
<td>2.5151</td>
</tr>
</tbody>
</table>

\( T - T_c = 0 \) as shown in figures 8.9 and 8.10.

\[
p^* = -1.2493(T^* - T_{c^*}) + 0.0274
\]

\( R^2 = 0.99903 \)

The critical pressure we obtain by this procedure is \( p_c = (0.0276 \pm 0.005) \). The same method was used to calculate the critical temperatures of SW fluids with \( \lambda \) equal to 1.5 and 2. The results are summarized in table 8.2.

We perform cluster distribution within the mesophase. Results support the colloidal nature of this region as shown in appendix E.

8.4 Short discussion

For the adhesive-sphere limit, the present value obtained for \( T_c \) by this procedure (0.2007) is consistent with that found by Largo and coworkers (Largo, Miller, and Sciortino, 2008). On the other hand, the critical pressure calculated is slightly higher than the one calculated
8.4. Short discussion

using subcritical two-phase coexistence measurements (0.54), of SW fluid with $\lambda = 1.005$ and $T_c = 0.2007$.

For $\lambda$ equal to 1.5, Largo and coworkers report a critical temperature of 1.0 using the optimized-cluster theory (Largo, Miller, and Sciortino, 2008). This seems somehow lower than the value calculated using our empirical method ($T_c = 1.18$). However, Mullen reported a critical temperature of 1.2 and 2.7 for $\lambda$ equal to 1.5 and 2 respectively (Mullen, 2012), using Wang-Landau simulations. Moreover, Lang et al. reported a critical temperature of 1.28 for $\lambda$ equal to 1.5 using the optimized random-phase approximation (Lang et al., 1999). These are in agreement with our results as shown in table 8.2.

We have introduced a new empirical method for calculating the critical parameters using only data from supercritical isotherms. The critical temperature calculated for SW fluids with $\lambda$ equal to 1.005, 1.5, and 2 are consistent with literature. This new empirical method presented is efficient and provides accurate calculations of critical temperature even for strongly short-ranged systems which are especially difficult to equilibrate.

Finally, we note that a colloidal nature of the supercritical mesophase is consistent with the well-known phenomenon of critical opalescence. The present interpretation of the thermodynamics of criticality provides an alternative description of critical opalescence observed in both molecular liquids and protein solutions, known as Tyndall scattering, which is a manifestation of a colloidal nature arising from light scattering by colloidal particles with sizes comparable to the wavelengths of white light. This method could then be applied to more realistic short-ranged potentials for proteins which could be patchy spheres or possibly non-spherical shapes interacting through anisotropic interactions. Identifying the critical temperature and how it depends on the pair potential is very important in formulations as we need to predict when the solution will go opalescent. For the adhesive-sphere limit, this linear mesophase region, where opalescence occurs due to density fluctuations extends to $T = 1.12T_c$. In protein solution, opalescence disappears at temperatures equal to 1.11 times the critical temperature as shown by Mason and coworkers (Mason et al., 2011), where the opalescence of an antibody solution changes significantly over a temperature range of 290 to 265 K or temperature ratio of 1.1 which is very close to our estimation.
Chapter 9

Project conclusions

The improving market for protein pharmaceuticals and the prevalence of aggregation-related diseases ensures that protein aggregation will continue to be an area of intense research. Molecular simulation serves as an important complement to experimental studies aimed at understanding and abating unwanted aggregation. Special attention is delivered to aggregation from the native state, mostly relevant to biopharmaceutics and bioprocessing, with a more sophisticated model (off-lattice bearing MJ potential) than the ones existing in literature addressing the same problem. Here we have focused our efforts on studying the correlation between low-temperature fluctuations and aggregation propensity of chains.

The event-driven molecular dynamics simulations of the off-lattice MJ heteropolymer model provided sufficient evidence that the model could capture the generic folding-characteristics of a protein. Model parameters can be fine-tuned to exhibit various folding behavior, whether it is a single-step folding for small fast-folding chains or multiple-step folding of more complex protein-chains. Chain length with 64 residues appears to be a short-enough chain capable of reproducing all characteristics of real proteins. By increasing the number of residues, transitions merge into a single transition from the extended state to the low-temperature state which is a behavior popular in protein chains. However, the computational intensity scales with the size of the chain making the simulation of large chains in multichain system timely inefficient. Moreover, although smaller chains appear computationally attractive, they lack the formation of a well-defined core at low temperature and hence native conformations are highly unstable. Therefore a 64-mer model appeared to be a good compromise between generic behavior of folding and computational intensity. Secondly, the overlap parameter influences the stability and the folding behavior of the model-chains. Increasing the
overlap parameter increases the number of transitions required for folding a chain into the low-T structure, but also increases the stability of the collapsed states.

The effect of increasing the range of interaction potential is similar to that found by increasing the overlap parameters: both increase the stability of the chains at high and low temperature, and increase the number of transitions needed for chains folding. Similarly by reducing both parameters, chains appeared less stable at high and low temperature, and more difficult to equilibrate. Increasing $\lambda$ is coupled with a shift of the melting and the collapse temperature (if different) to higher temperatures indicating an increase in stability of the collapsed states. Moreover, a critical value for the range of interaction potential above and below which the chains fold in different ways. This value is not very exact as the difference in $\lambda$ between chains is in the order of 0.2. For $\lambda \geq 1.3$, all chains fold through multiple transitions having the collapse and the freezing transitions common between them. By decreasing $\lambda$, intermediate transitions (transition other than the collapse and freezing transitions) merge first, then the collapse and the freezing transitions come together and chains fold in a single step transition with all-or-none characteristics as shown by looking at chains with $\lambda \leq 1.1$. Accordingly, the model chosen is a 64 monomer chain with $\lambda = 1.1$, and $\sigma/l = 1.9$ appeared the least computationally expensive, displaying all the characteristics found in real proteins. These parameters represent the smallest chain with the limit in the range of interaction below which chains fold in a single step, i.e. $\lambda = 1.1$, with large stiffness $\sigma/l = 1.9$ to maximize the stability of the native structure at low temperature, and its resistance to unfolding at high temperature.

In the next section, we found an order parameter that allows us to distinguish between folded and unfolded states. This allows us to distinguish fluctuations within the low-temperature structures or equivalently breathing motion (contact and conformational fluctuations) from partial unfolding of the collapsed states. Motivated by the role of the core in the folding and aggregation of proteins, the order parameter was defined as the average root mean square in core contacts $<\text{RMSC}>$, with a critical value of 0.8, where chains with an $<\text{RMSC}>$ above the critical value are considered to be folded. This corresponds to a value in $<\text{RMSD}_{CM}>$ equal to 0.7 (occurring at the same temperature). This order parameter was validated as this critical value occurs at the melting temperature of the chains simulated.
which in turn corresponds to the freezing transition in most models and to a higher temperature transition for unusually stable structures with exceptionally larger core. Also, the stability of the low-T was correlated to the stability of high-T states and it showed that some chains with more stable low-T ensemble correlates with unstable collapsed states.

Finally, the behavior of the chains is similar to that of chains in isolation and it is possible to extrapolate the aggregation propensity of the model chains from behavior at room temperature of isolated chains. Based on the data of $<\text{RMSC}>$ and $<E_{\text{inter}}>$, conformational fluctuations at temperatures below the unfolding/melting temperature do not induce aggregation and are correlated with breathing motion of the chains. At higher temperature unfolding induces interchain contacts irreversibly. Therefore partially unfolded intermediates are the main precursor of aggregation.
Chapter 10

Recommendations for future work

This work is merely a stepping stone in the field of proteins simulations. Since this work has introduced a new property as a divide for the structures between folded and unfolded, it is only reasonable to revisit previous work in the light of this definition. Moreover, since the core is shown to play a major effect in the aggregation propensity one option is to investigate the effect of point mutations of the beads buried inside the chain and forming the core. This gives insight on the relative sensitivity of the core to interaction change and how strong does it hold the chain together. Hopefully, using statistical analysis of various chains it would be interesting to define a relative strength threshold below which the core have low resistance to temperature increase and unfolds easily.

On another note, multichain system for 64-mer sequences are simulated up to 8 chains. However, the computational power of DynamO allows the simulation up to hundreds of chains. Given that we have many systems with different lengths, range of interactions, and stiffness values, the effect of those parameters on the aggregation propensity can be investigated. The effect of the box size (which simulates the concentration) can also be investigated. These large systems are obviously a more realistic system, as proteins exist in ‘real’ solutions in thousands.

Finally, since our model can be fine-tuned to exhibit various folding pathways, many proteins well-investigated in experiments can be simulated if the intermediate formed are well-known. Once this happens, simulation can give insight on the aggregation phenomena of these specific proteins, inaccessible to conventional experiments.

Ultimately, mapping all simulation parameters to reality would concretize and quantify the
simulation results. This includes mapping low temperature, melting temperature, range of interaction potential, stiffness, ... etc, to reality. Moreover, mapping the solution pH, protein concentration, denaturant, chaperons, etc, to simulation would contribute largely to our fundamental understanding of the phenomena being analyzed.
Appendix A

Effect of interaction potential on a 64monomer chain with an overlap parameter of 1.3

Figures A.2 and A.1 are plots of variation of heat capacity and radius of gyration respectively for several 64monomers having different range of interactions.

Similarly to the chains with stiffness of 1.3, the trend is conserved: Chains with long-ranged interaction (1.5 and 1.3 $\lambda$) undergoes a two-step folding. This is apparent in the multiple peaks in the $C_v$ plot, and the continuous decrease in the range of gyration. By reducing the range of interactions, the two transitions come together and the chains fold in a first-order discontinuous transition. This is shown in the sharp peak in the $C_v$ plot and the steep slope in the $R_g$ plot, indicating a sudden collapse in the size of the chain associated with this transition into the ground state. It should be noted that by reducing $\lambda$, transitions shift to lower temperatures, indicating a destabilization of the structure.
Effect of range of interaction potential on the folding of 64mer

Stiffness, $\sigma = 1.3$

Figure A.1: A plot that shows the differences in the specific heat capacity as a result of varying the range interaction potential.
Appendix A. Effect of interaction potential on a 64monomer chain with an overlap parameter of 1.3

Effect of range of interaction potential on the folding of the 64mer

Stiffness of the chain, $\sigma = 1.3$

**Figure A.2**: A plot of the radius of gyration in function of temperature for six 64-mer chains with different range interaction potential and stiffness $\sigma = 1.3$. 

Dimensionless temperature, $T^*$

Radius of gyration, $R_g^*$
Appendix B

Isolated chains low-temperature structure

We present plots of the temperature dependence of the specific heat capacity B.1(a), radius of gyration B.1(b), and reweighted internal energy B.1(c) in figure B.1. This figure will be revised for in-depth analysis. For all chains simulated, the increase in heat capacities coincides with a decrease in reweighted internal energy and a reduction in the size of the chains. This occurs at $T^* = 0.31$ for chains 1, 2, 3, and 5, and at a higher temperature of at $T^* = 0.39$ for chain 4. Although it is expected to have a single dominant peak in $C_v$, but the use of large overlap parameter introduces high stiffness to the chains associated with a continuous conformational change and a step-wise bending from the extended coil to the compact structure, by the means of intermediate structure(s). This possibly explains the presence of smaller peak surrounding the main one. These small transitions are not expected to have a thermodynamic relevance to the folding behavior.

Figure B.2 shows the average contact maps at the low-temperature $T^* = 0.033$ of chains 1, 2, 3 and 4. Another advantage of using square well potentials is that contact maps can be used to visualize and unambiguously define a protein structure in 2-dimensional space. This also allows to define states (i.e. native versus non-native) and calculate free energy differences between those states. The contact map is almost entirely formed of black and white squares. This indicates that contacts between residues are permanent and well-defined. This further characterize the existence of a single structure at this temperature. The appearance of very few gray shades can be correlated to the ‘breathing-motion’ of real proteins at low-temperatures as explained previously.
The main characteristic of a protein model-chain is to form a ‘native state’ at low temperature. Care is taken when using the term ‘native-state’ here, as we will introduce in the following section a new view of the low-temperature conformation and the existence of rather a native-ensemble as a broad basin at the bottom of the energy funneled-landscape.

In figure B.3 we present histogram plots of probabilities of observing a cluster of structures
Appendix B. Isolated chains low-temperature structure

Figure B.2: Average contact map at the lowest temperature $T^* = 0.036$ for chain 1 (a), chain 2 (b), chain 3 (c) and chain 4 (d). This is an 64x64 matrix, where 64 is the number of beads in the chain. This matrix is plotted in a 2D scheme showing the contacts between any pair of monomers. The monomers are presented on the two axis, in correlation with their position in the chain. Boxes represent the probability of occurrence for each contact. A black square denotes that no contact in made, a white square shows that the pairs are in full contact. A gray square suggests that the pair of monomer is in partial contact, with the intensity of the color indicating how often or how probable is the contact made. Contact maps are generated for each replica and gives access to the configurations at each stage of the stage of the process.

with a given deviation in conformation (a), in contacts (b), and in contacts of the core (c).

The RMSD histogram plot in figure B.3 (a) with a cut off equal to 0.5 exhibit a single peak in histogram (chains 2, 3, and 4) slightly above zero indicating that the conformation of all structures at the lowest temperature are fluctuating within a deviation less or equal to 0.5 and can therefore be clustered into a single cluster. This slight shift in RMSD can be related to the low-temperature breathing of the chains. The RMSD histogram of chain 1 appears more shifted to higher RMSD values around 0.3.

Figures B.3 (b) and B.3 (c) are histograms of probability of the contacts of the chain and contacts of the core respectively. The histograms are well-defined and narrow peaks at RMSD$_{CM}$ and RMSC equals to 1 indicating the existence of specific stable contacts in the chain and between buried residues. The chains therefore exhibit a unique single conformation at the lowest temperature with stable core.
Figure B.3: Probability histograms at the lowest temperature $T^* = 0.033$ of (a) The average root mean square deviation $rmsd_{a,b}$, (b) Average root mean square deviation of contacts $rmsdc_{a,b}$, and (c) the average root mean square deviation of contacts of the core.

In this and the following analysis, we showed that our model-chain exhibit generic behavior of ‘real’ proteins. Chains undergo a collapse from a high-temperature extended-coil to
a low-temperature compact structure. This low-energy structure exhibit a stable buried hydrophobic core and minor conformational fluctuations. This shows that our model is a good model for studying protein aggregation.
Appendix C

Temperature lists
### Table C.1: Temperature list of the three 64-mer with $\lambda$ equal to 1.1 and $\sigma/l$ equal to 1.9, 1.6, and 1.3; and four 64-mer with $\lambda$ equal to 1.5 and $\sigma/l$ equal to 1.9, 1.6, 1.5, and 1.0.

<table>
<thead>
<tr>
<th>Chains</th>
<th>$\lambda = 1.1$</th>
<th></th>
<th></th>
<th>$\lambda = 1.5$</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>$\sigma/l = 1.3$</td>
<td>$\sigma/l = 1.6$</td>
<td>$\sigma/l = 1.9$</td>
<td>$\sigma/l = 1.0$</td>
<td>$\sigma/l = 1.3$</td>
<td>$\sigma/l = 1.6$</td>
</tr>
<tr>
<td>0</td>
<td>0.0394</td>
<td>0.0391</td>
<td>0.033</td>
<td>0.037</td>
<td>0.049</td>
<td>0.046</td>
</tr>
<tr>
<td>1</td>
<td>0.0399</td>
<td>0.0395</td>
<td>0.034</td>
<td>0.043</td>
<td>0.057</td>
<td>0.053</td>
</tr>
<tr>
<td>2</td>
<td>0.0405</td>
<td>0.04</td>
<td>0.035</td>
<td>0.051</td>
<td>0.067</td>
<td>0.062</td>
</tr>
<tr>
<td>3</td>
<td>0.041</td>
<td>0.0405</td>
<td>0.036</td>
<td>0.059</td>
<td>0.083</td>
<td>0.071</td>
</tr>
<tr>
<td>4</td>
<td>0.0425</td>
<td>0.041</td>
<td>0.0367</td>
<td>0.067</td>
<td>0.099</td>
<td>0.083</td>
</tr>
<tr>
<td>5</td>
<td>0.044</td>
<td>0.0425</td>
<td>0.0375</td>
<td>0.083</td>
<td>0.122</td>
<td>0.099</td>
</tr>
<tr>
<td>6</td>
<td>0.046</td>
<td>0.044</td>
<td>0.039</td>
<td>0.099</td>
<td>0.144</td>
<td>0.122</td>
</tr>
<tr>
<td>7</td>
<td>0.048</td>
<td>0.046</td>
<td>0.043</td>
<td>0.112</td>
<td>0.168</td>
<td>0.144</td>
</tr>
<tr>
<td>8</td>
<td>0.052</td>
<td>0.048</td>
<td>0.05</td>
<td>0.128</td>
<td>0.192</td>
<td>0.178</td>
</tr>
<tr>
<td>9</td>
<td>0.057</td>
<td>0.052</td>
<td>0.055</td>
<td>0.144</td>
<td>0.225</td>
<td>0.212</td>
</tr>
<tr>
<td>10</td>
<td>0.06</td>
<td>0.057</td>
<td>0.068</td>
<td>0.166</td>
<td>0.261</td>
<td>0.261</td>
</tr>
<tr>
<td>11</td>
<td>0.066</td>
<td>0.06</td>
<td>0.085</td>
<td>0.189</td>
<td>0.31</td>
<td>0.31</td>
</tr>
<tr>
<td>12</td>
<td>0.07</td>
<td>0.066</td>
<td>0.1</td>
<td>0.212</td>
<td>0.383</td>
<td>0.383</td>
</tr>
<tr>
<td>13</td>
<td>0.074</td>
<td>0.07</td>
<td>0.12</td>
<td>0.261</td>
<td>0.455</td>
<td>0.455</td>
</tr>
<tr>
<td>14</td>
<td>0.078</td>
<td>0.074</td>
<td>0.13</td>
<td>0.3</td>
<td>0.556</td>
<td>0.556</td>
</tr>
<tr>
<td>15</td>
<td>0.082</td>
<td>0.078</td>
<td>0.15</td>
<td>0.34</td>
<td>0.9</td>
<td>0.666</td>
</tr>
<tr>
<td>16</td>
<td>0.086</td>
<td>0.082</td>
<td>0.16</td>
<td>0.383</td>
<td>2.092</td>
<td>2.092</td>
</tr>
<tr>
<td>17</td>
<td>0.09</td>
<td>0.086</td>
<td>0.168</td>
<td>0.444</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>0.098</td>
<td>0.09</td>
<td>0.176</td>
<td>0.556</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>0.105</td>
<td>0.098</td>
<td>0.184</td>
<td>0.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.11</td>
<td>0.105</td>
<td>0.2</td>
<td>2.092</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>0.119</td>
<td>0.11</td>
<td>0.225</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>0.122</td>
<td>0.119</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>0.128</td>
<td>0.122</td>
<td>0.31</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0.135</td>
<td>0.128</td>
<td>0.39</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.142</td>
<td>0.135</td>
<td>0.47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>0.146</td>
<td>0.142</td>
<td>0.58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>0.152</td>
<td>0.146</td>
<td>2.092</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>0.1585</td>
<td>0.152</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>0.165</td>
<td>0.1585</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>0.172</td>
<td>0.165</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>0.182</td>
<td>0.172</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>0.188</td>
<td>0.182</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>0.195</td>
<td>0.188</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>0.201</td>
<td>0.195</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>0.208</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>0.212</td>
<td>0.206</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>0.225</td>
<td>0.212</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>0.24</td>
<td>0.225</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>0.25</td>
<td>0.24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>0.26</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>0.29</td>
<td>0.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>0.32</td>
<td>0.32</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>0.4</td>
<td>0.461</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>0.461</td>
<td>0.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>1</td>
<td>2.092</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>2.092</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix D

Multichain system, a detailed analysis

Figure D.1 show the temperature dependence of intermolecular energy for 5 systems each containing 8 copies of the same peptide, where the starting configuration corresponds to the representative structure obtained from clustering the structures at a temperature of 0.033.

FIGURE D.1: Temperature dependence of intermolecular for multichain system of 8 64-monomer chains replicated from the native state of (a) chain 1, (b) chain 2, (c) chain 3, (d) chain 4, and (e) chain 5.
For system 1, at temperature lower than 0.16, $E_{inter}$ is almost constant. The interchain energy exhibit major fluctuations at temperature equal to 0.16 and 0.18 which corresponds to the freezing and the collapse temperature of the chain in isolation as shown in the figure 6.2. In fact for the isolated chain the freezing transition is a disordered-ordered transition indicating that structures below the freezing temperature are disordered and do not take part of the native ensemble. So it was expected that these structures are more prone to aggregation. This is consistent with the data in multichain systems as the interaction between chains started to increase exponentially for temperatures above 0.16. At temperature equal to 0.18, there is another change in the behavior of $E_{inter}$ where it slightly increases then drops towards a minimum. This fluctuation occurs at the temperature of the collapse transition.

The behavior of the chains in system 2 is different from the other systems. The fluctuations in interchain energy appear to propagate from the lowest temperature equal to 0.05 to the highest temperature. Effectively, the increase in interchain energy at 0.05 appears to cause aggregation because it does not stabilize at higher temperature, rather fluctuations gradually increase across the temperature range. This temperature corresponds to the freezing temperature of chain 2 as shown in the previous section.

For system 3, the interchain energy is constant between 0.33 and 0.14. For temperature between 0.14 and 0.2, $E_{inter}$ unexpectedly increases on average, and at temperature equal to 0.2 it drops rapidly towards the minimum value in interchain energy. By mapping this behavior to folding of the single chain in isolation, we find that temperature equal to 0.14 corresponds to a freezing order-order transition. Above this temperature, the structures are still considered ordered/folded with an <RMSC> above its critical value. The temperature equal to 0.2 corresponds to the peak of the collapse transition in heat capacity of the isolated chain at which the <RMSC> is at the critical value as shown in figure 6.5. Above this temperature structures are considered disordered or unfolded. This is consistent with the data in multichain system in figure D.1 (c) where the interchain energy exhibit the largest decrease at temperature equal to 0.2 indicating aggregation or interchain contacts.

For system 4, the interchain energy gradually fluctuates between the lowest temperature and the temperature equal to 0.16. At higher temperatures, $E_{inter}$ decreases with a near infinite slope towards minimum. This indicates that structures at temperatures higher than 0.16 interact and aggregate. The temperature equal to 0.16 corresponds to the disorder-order
freezing transition of the isolated chain 4 and overlaps with the critical temperature where the $\langle \text{RMSC} \rangle$ is equal to 0.8 as shown in figure 6.3. When studying the single chain system we argued that chains at temperatures higher than the freezing transition (equivalently with $\langle \text{RMSC} \rangle$ lower than 0.8) does not take part of the native ensemble and are prone to aggregation.

For system 5, the interchain energy fluctuates slightly between the lowest temperature and 0.18. Above this temperature $E_{\text{inter}}$ decreases towards a minimum across a small temperature range. The temperature of 0.18 corresponds to the freezing disordered-ordered transition of the chain in isolation below which the structures are in the native ensemble. This is also the temperature at which the $\langle \text{RMSC} \rangle$ of the single chain is at the critical value. This indicates that fluctuations within the native ensemble do not induce aggregation, while fluctuations above the freezing transition promote aggregation.

To illustrate the relationship between the stability of the core and the formation of interchain contacts, we present in figure D.2 the deviation in the root-mean-square of contacts of the core as a function of reduced temperature for the five systems. In all the system, at the lowest temperature $T^* = 0.05$, all chains have an RMSC approximately equal to one. As the temperature is increase to $T^* = 0.16$ or below, chains in the 5 systems have an RMSC between 0.8 and 0.9. At $T^* = 0.184$, deviations in the RMSC becomes more distinct. In the first system, only 2 chains have an RMSC of 0.8 and above, while the RMSC of the 6 other chains is in the rang of 0.55 to 0.65. The average RMSC at $T^* = 0.184$ is equal to 0.68. In the second system, 4 chains have an RMSC between 0.88 and 0.95, 3 chains have a slightly lower RMSC between 0.7 and 0.8. One chain have a surprisingly low RMSC of 0.53. The average RMSC at $T^* = 0.184$ is 0.8. In the third system, one chain have a high RMSC of 0.96, 7 chains have an RMSC around 0.80, with an average RMSC of 0.82. In the fourth system, one chain has an RMSC of 0.9, 4 chains have an RMSC between 0.8 and 0.85. The 2 remaining chains have an RMSC between 0.75 and 0.8, with a total average RMSC of 0.8. System 5 shows a close behavior to the first system, with an equal average RMSC of 0.68. Two chains have an RMSC between 0.8 and 0.9. Two other chains have an RMSC around 0.7. The remaining peptides exhibit a low RMSC of 0.6 and below.

At $T^* = 0.21$, Systems 1, 2, and 3 have a an average RMSC of the chains around 0.55. The collapsed states in chains in system 4 appear to be more stable with a total average RMSC
Figure D.2: Temperature dependence of root mean square deviation in contacts of the core for multichain system of 8 64-monomer chain replicated from the native state of (a) chain 1, (b) chain 2, (c) chain 3, (d) chain 4, and (e) chain 5.

Of 0.65. In system 5 on the other hand the collapsed states exhibit the least stability with the highest average RMSC of 0.35.
Appendix E

Cluster Distribution in SW fluids

E.0.0.1 Cluster Distribution

MD configurations along the slightly supercritical isotherm ($T^* = 0.215$) have been carried out to determine the probability distribution of finding a particle in a cluster of size $n$ at equilibrium. The use of SW fluids with short-range attraction makes cluster counting unambiguous were defined a bond as a pair of particles with relative distance less than $\sigma + \lambda$. At very low density, i.e. for an ideal gas, there are no clusters and

$$\frac{n_c}{N} = 1$$ (E.1)

where $n_c$ is the number of clusters. As the density increases, there are an ever-increasing number of pairs, then triplets etc., and increasingly large molecular clusters, until reaching a density when just one cluster of size $N$ fills the whole box. At this point the probability function $(n_c - 1)/N$ goes to zero for large $N$. The cluster distribution for the intermediate density $\rho^* = 0.36$, i.e. within the region of the mesophase, is shown in figure E.1.

Figure E.1 shows that there is clear mixture of gas phase and liquid phase in the supercritical mesophase. The gas component at the density 0.36 is characterized by a distribution of molecular sizes, monomers (12%) dimers (6%), trimers (4%), etc. with ever-decreasing probability, with an extremely small probability extending up to several hundred. The liquid phase, by contrast, shows just one cluster that is of the order of the size of the system, but variations in size in probability result as gas and liquid fractions fluctuate. At high temperatures, with $T = 1.2T_c$, this cluster dist The data in figure 8.7 are indicative of a linear
Appendix E. Cluster Distribution in SW fluids

Figure E.1: Distribution of clusters at the density $\rho^* = 0.36$ for a supercritical isotherm at temperature $T^* = 0.215$: the probability plot shows that all the atoms in the system belong either to a gas phase (molecules of less than 400) or liquid phase with clusters of the order of the system size $> 5000$; the hiatus shows there is no intermediate cluster sizes.

The combination rule for thermodynamic state functions, which explains the linear equation-of-state for the pressure $p(\rho)T$ in the supercritical mesophase, as noted previously, and seen here in figure 8.7, for example.

In figure E.2 we show the ensemble equilibrium average size of a cluster as a function of density: the mesophase is characterized by a steep sinusoidal increase in the cluster size on increasing the density as the nature of the colloidal mesophase changes from liquid-droplets in gas to gas micro bubbles (voids) in liquid.

Another more illuminating way to reveal the colloidal nature of the mesophase is by computer graphics, using color-coding for clusters of different sizes. In figure E.3, we show pictures of the molecular structures along a supercritical isotherm for densities corresponding to the gas phase (E.3 a), the mesophase, on the gas side (E.3 b) the mesophase on the liquid side (E.3 c), and liquid phase (E.3 d). Figure E.3 (a) shows the gas molecules with a predominance of monomers (red) and all other sized molecules (brown); the largest molecules are around 30 atoms at this density, albeit with very low probability. Figure E.3 (b) represents an equilibrium configuration within the mesophase at the density $\rho^* = 0.36$. The
color-coding for the gas phase is the same as in figure E.3 (a), except that the largest gas cluster, 267 atoms, is shown as yellow. There are two very large liquid clusters in this particular snapshot colored dark blue (4042 atoms) and second cluster colored light blue (1129). The important point to note is that every atom or molecule can be designated liquid or gas phase. There are no clusters of an intermediate size, i.e. there is no “continuity of gaseous and liquid phases”. We see a colloidal mixture that is nanoscopically heterogeneous but thermodynamically homogeneous, as distinct from the subcritical coexistence region that is thermodynamically heterogeneous. In figure E.3 (c), on the liquid side of the mesophase, \( \rho^* = 0.4 \), the mixture of gas molecules and liquid cluster is also evident. The main difference between figure E.3 (b) and figure E.3 (c) is the gas side of the mesophase (b), the dispersed phase is the liquid nanodroplets, (resembling a colloidal mist) whereas on the liquid side of the mesophase (c) the dispersed phase is the nano-bubbles of gas (resembling a colloidal foam). Finally, for densities above the critical percolation threshold, as seen in figure E.3 (d), there is just one large liquid cluster with a very dilute solution of gas like monomers or dimers.
FIGURE E.3: Snapshots of instantaneous configurations of the model fluid along a supercritical isotherm $T^* = 0.215 (T/T_c = 1.075)$ Atoms as gas phase monomers are red, and molecules ($n \geq 2$) brown; all atoms in liquid state clusters are blue; in the mesophase ((b) and (c)), the largest clusters of gas molecules are green and yellow. (a) gas phase $\rho^* = 0.2$; (b) mesophase gas-side $\rho^* = 0.36$; (c) mesophase liquid-side $\rho^* = 0.4$; (d) liquid phase $\rho^* = 0.7$. 
Bibliography


Booth, D. R. et al. (1997). “Instability, unfolding and aggregation of human lysozyme variants underlying amyloid fibrillogenesis”. In: NATURE 385.6619, pp. 787–793. ISSN: 0028-0836. DOI: 10.1038/385787a0.


BIBLIOGRAPHY


Chiti, F. et al. (2002). “Kinetic partitioning of protein folding and aggregation”. In: NATURE STRUCTURAL BIOLOGY 9.2, pp. 137–143. ISSN: 1072-8368. DOI: 10.1038/nsb752.


Ding, F. et al. (2002). “Molecular dynamics simulation of the SH3 domain aggregation suggests a generic amyloidogenesis mechanism”. In: JOURNAL OF MOLECULAR BIOLOGY 324.4, pp. 851–857. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(02)01112-9.


Hall, Carol K. and Vcroria A. Waggner (2006). “Computational approaches to fibril structure and formation”. In: Amyloid, Prions, and Other Protein Aggregates, Pt B 412, pp. 338–365. DOI: 10.1016/S0076-6879(06)12020-0.


Hammarstrom, P. et al. (1999). “Structural mapping of an aggregation nucleation site in a molten globule intermediate”. In: JOURNAL OF BIOLOGICAL CHEMISTRY 274.46, pp. 32897–32903. ISSN: 0021-9258. DOI: 10.1074/jbc.274.46.32897.


Mullen, Ryan Gotchy (2012). “Phase Equilibria for Lennard-Jones and Square-Well Fluids”. In: Journal of Chemical Thermodynamics.


– (2006). “Spontaneous fibril formation by polyalanines; Discontinuous molecular dynamics simulations”. In: Journal of the American Chemical Society 128.6, pp. 1890–1901. DOI: 10.1021/ja0539140.


– (1996). “Simulations of reversible protein aggregate and crystal structure”. In: BIOPHYSICAL JOURNAL 70.6, pp. 2888–2902. ISSN: 0006-3495.


Safar, J. et al. (1994). “Scrapie amyloid (prion) protein has the conformational characteristics of an aggregated molten globule folding intermediate”. In: BIOCHEMISTRY 33.27, pp. 8375–8383. ISSN: 0006-2960. DOI: 10.1021/bi00193a027.


BIBLIOGRAPHY


Yoshii, H. et al. (2000). “Refolding of denatured/reduced lysozyme at high concentration with diafiltration”. In: BIOSCIENCE BIOTECHNOLOGY AND BIOCHEMISTRY 64.6, pp. 1159–1165. ISSN: 0916-8451. DOI: 10.1271/bbb.64.1159.
