

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

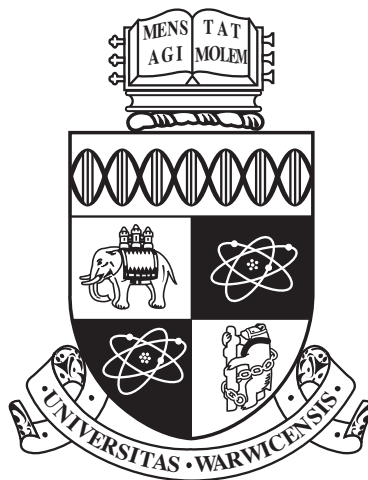
A Thesis Submitted for the Degree of PhD at the University of Warwick

<http://go.warwick.ac.uk/wrap/74165>

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.



**Coarse-Grained Simulations of Intrinsically
Disordered Peptides**

by

Gil Rutter

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Physics

October 2015

THE UNIVERSITY OF
WARWICK

Contents

List of Tables	iv
List of Figures	vi
Acknowledgments	xvii
Declarations	xviii
Abstract	xix
Chapter 1 Introduction	1
1.1 Protein structure	1
1.1.1 Dependence on ambient conditions	3
1.2 Protein disorder	6
1.2.1 Determinants of protein disorder	6
1.2.2 Functions of IDPs	7
1.3 Experimental techniques	8
1.3.1 Intrinsic disorder	8
1.3.2 Biomineralisation systems	10
1.4 Molecular dynamics for proteins	12
1.4.1 Approaches to molecular simulation	12
1.4.2 Statistical ensembles	13
1.5 Biomineralisation	15
1.6 n16N	16
1.7 Summary	20
Chapter 2 Accelerated simulation techniques	21
2.1 Coarse-graining	21
2.1.1 Defining the CG mapping	22
2.1.2 Explicit and implicit solvation	24

2.1.3	Validating a CG model	26
2.1.4	Choices of coarse-grained models	26
2.2	Accelerated sampling techniques	29
2.2.1	Replica exchange molecular dynamics	30
2.2.2	Statistical temperature molecular dynamics	31
2.3	Summary	34
Chapter 3	Software modifications and validation	35
3.1	PRIME20 initial parametrisation	35
3.1.1	PRIME model validation	36
3.1.2	Parametrising a PRIME20-like model	40
3.2	Implementation work in LAMMPS	42
3.2.1	The PLUM model	42
3.2.2	Statistical temperature molecular dynamics	46
3.3	Summary	52
Chapter 4	Single-chain simulations	53
4.1	PLUM model simulations	53
4.1.1	Over-stabilisation of the α -helix	53
4.1.2	PLUM* validation work	54
4.1.3	The n16N-1 system in PLUM*	58
4.1.4	The n16NN-1 system	61
4.1.5	The S1 system	62
4.2	PRIME20-like model simulations	64
4.2.1	The S1 system	64
4.2.2	The n16N-1 system	65
4.2.3	The n16NN-1 system	70
4.3	Summary	71
Chapter 5	Multiple-chain simulations	73
5.1	The n16N-2 system	73
5.1.1	PLUM*	73
5.1.2	PLUM	74
5.1.3	PRIME20-like	78
5.2	The n16NN-2 system	81
5.2.1	PLUM*	81
5.2.2	PRIME20-like	82
5.3	Discussion of dimer systems	83

5.4	The n16N-3 system	84
5.4.1	PLUM*	84
5.4.2	PRIME20-like	85
5.5	The n16NN-3 system	87
5.6	Discussion of trimer systems	90
5.7	The n16N-6 system	91
5.7.1	PLUM*	91
5.7.2	PRIME20-like	92
5.8	The n16NN-6 system	97
5.9	Discussion of hexamer systems	99
5.10	Summary	101
Chapter 6 Conclusions		103
6.1	PLUM* model	103
6.1.1	Simulation results	103
6.1.2	Further work	105
6.2	The PRIME20-like model	106
6.2.1	Simulation results	106
6.2.2	Further work	107
6.3	Outlook for coarse-grained models to study IDPs	108

List of Tables

1.1	<i>Relative frequencies of residues in common secondary structure motifs [Creighton, 1992] [Berg et al., 2002, page 67] and disorder propensity according to the TOP-IDP scale [Campen et al., 2008], in which higher values are more disordered. This is one of many scales on which residue disorder propensity has been ranked [Galzitskaya et al., 2006; Oldfield et al., 2005; Vihinen et al., 1994; Garbuzynskiy et al., 2004]. The highest values are highlighted.</i>	7
1.2	<i>Summary of experimental research on n16N peptide, with a focus on experimental techniques used. All assays contain n16N in water solution with calcium and carbonate ions for crystal growth.</i>	
	<i>AFM - Atomic force microscopy</i>	
	<i>CD - Circular dichroism spectroscopy</i>	
	<i>DLS - Dynamic light scattering</i>	
	<i>EDX - Energy dispersive X-ray spectroscopy</i>	
	<i>FM - Fluorescence microscopy</i>	
	<i>FTIR - Fourier-transformed infrared spectroscopy</i>	
	<i>Raman - Raman spectroscopy</i>	
	<i>SEM - Scanning electron microscopy</i>	
	<i>TEM - Tunnelling electron microscopy</i>	
	<i>XANES - X-ray absorption near edge spectromicroscopy</i>	
	<i>XRD - X-ray diffraction</i>	
	<i>X-PEEM - X-ray photoelectron emission spectromicroscopy</i>	11
1.3	<i>Suggested roles of the subdomains of n16N [Brown et al., 2014]. . .</i>	20

4.1 *Top interactions between residues ranked by frequency of occurrence. To qualify, any atom on residue A has to be in the square well of any atom on residue B, with A and B separated with at least three residues in-between. Tyrosine dominates the rankings, and is clearly a cornerstone of intrapeptide stabilisation.* 69

List of Figures





- 1.1 Wikimedia user Dcrjsr, available under the Creative Commons Attribution 3.0 Unported license.

The backbone atoms of a central amino acid residue is shown, with two neighbours labelled (-1) and (+1) partially shown. Each residue except glycine also has a side-chain, denoted $C\beta$. The dihedral angles ϕ , ψ and ω which determine the secondary structure are labelled. . 2

- 1.2 [Hollingsworth and Karplus, 2010] adapted from [Ramachandran and Sasisekharan, 1968].

A Ramachandran plot showing both example data and outlines of commonly accepted regions. The data is comprised of 63,149 residues from crystal structures and each point indicates the (ϕ, ψ) values representing the secondary structure of a single residue.

The outlines are divided into core allowed regions (solid lines) and allowed regions (dashed lines). Several forms of secondary structure have their location indicated; these are α -helix (α), 3_{10} -helix (3), π -helix (π), left-handed α -helix (α_l), 2.2_7 ribbon (2), polyproline-II (II), collagen (C), parallel β -sheet ($\uparrow\uparrow$) and anti-parallel β -sheet ($\uparrow\downarrow$). . 4

1.3	<p><i>A steric map, in which steric clashes leading to forbidden regions are shown in (ϕ, ψ) phase space and labelled according to the clashing pair of atoms [Ho et al., 2003]. Dashed blue lines and blue labels denote clashes. Favourable dipole-dipole backbone interactions are also plotted, with red dashed lines and red labels. Areas of the map are white if forbidden, and otherwise coloured according to the legend below. The atom labels match the illustration in fig. 1.1.</i></p> <p>Boundaries</p> <p>- - - Attractive dipole-dipole interactions</p> <p>- - - Steric clashes</p> <p>Accessible regions</p> <p> Sterically permitted</p> <p> Single steric clash (outlier region)</p> <p> Left- and right-handed α-helix regions</p> <p> β-strand region</p>	5
1.4	<p>[Levi-Kalisman et al., 2001]</p> <p><i>A putative scheme for the organic biomineralisation matrix which grows nacre. Interlamellar matrix sheets are composed mainly of aligned β-chitin fibres in several layers, with acidic glycoproteins at their surface which lead to electron-dense patches (not labelled). Aragonite biomineralisation occurs in the disordered silk fibroin gel region, most likely nucleating epitaxially on the chitin framework [Weiner et al., 1984].</i></p>	17
1.5	<p><i>Amino acid sequence of the 30AA N-terminal region of n16, called n16N. An ellipsis indicates where the full n16 sequence continues, and braces indicate suggested subdomains [Brown et al., 2014], summarised in table 1.3. Cationic amino acid residues shown in bold blue, anionic residues shown in bold red. The last 14 residues, labelled SD3, represent a highly charged region which may be the mineral assembly subdomain.</i></p>	18

1.6 [Amos et al., 2011]

Two-stage mineralisation experiment to test whether pre-formed n16N assemblies could nucleate calcium carbonate polymorphs. In stage one, n16N deposits are allowed to form in typical mineralisation conditions (50 μ M n16N, 16h, 16 $^{\circ}$ C). In stage two, washed supports with n16N deposits are transferred to a mineralisation solution with or without n16N content.

(A) *Calcium carbonate mineralisation without n16N IDP. Calcite polymorph dominates.*

(B) *n16N present during first stage of mineralisation, yielding more vaterite (v).*

(C,D) *n16N present during both stages of mineralisation. The arrows indicate fibril-spheroidal deposits of n16N on the exposed surface of a crystal. 19*

2.1 *The PRIME20 and PLUM models are identical in their mappings of atoms to coarse-grained sites. All residues feature a single side-chain site, except for glycine which has no side-chain site. The PLUM model backbone maintains its spatial arrangement through continuous-potential bonds, angles and dihedrals as in equation (2.2), while the PRIME20 model relies upon square bonds and square pseudo-bonds, seen in fig. 2.1b. 27*

3.1 *Ramachandran plots showing the exploration of (ϕ, ψ) phase space for PRIME2001 chains, according to the reference [Voegler Smith and Hall, 2001], our independent reproduction, and a reproduction with a modified parameter set. Note that $T = 0.15$ is a high simulation temperature, certainly above that used by the original authors. By modifying the parameter set, it is possible to reproduce the reference data's accessible regions with moderate accuracy. 37*

3.2 *Behaviour of a chain of A20 in PRIME2001, PRIME2001-like and PRIME20-like models. After altering the 2001 model to reproduce the authors' accessible areas (see fig. 3.1), we see here that the hydrogen bond interaction strength, ϵ_{HB} , needs raising to 1.26 to match the reference data best. In the PRIME20-like model, α -helices are far more stable. Unfortunately, no data for the behaviour of A20 in the canonical PRIME20 model is available. 39*

3.3	<i>Ramachandran plots in the custom PRIME2001-like model at $T = 0.125$, with $\epsilon_{HB} = 1.26$. These plots illustrate that the PRIME2001 data in fig. 3.1a and 3.1b can be reproduced by the custom model.</i>	39
3.4	<i>Final structures of the 48-peptide $A\beta_{16-22}$ system in the PRIME20-like model. Both structures form β-sheets, with a preference for anti-parallel β-strands. Left: $T^* = 0.17$ simulation. Five sheets feature all 48 chains. Right: $T^* = 0.20$ simulation. The 45 chains shown make up four sheets, while one chain is free and two are disordered on the surface of the structure.</i>	43
3.5	<i>Validation data for the bonded components of the PLUM model, contrasted with the reference data [Bereau and Deserno, 2009]. The free energy landscape of the Ramachandran plot is shown for the GAG tripeptide at $T^* = 1.0$. The colour represents the free energy difference with the lowest conformation in reduced units. Both datasets result from REMD simulations at reduced temperatures 0.5, 0.7, 1.0, 1.3, 1.6, 1.9, 2.2 and 2.5. WHAM [Ferrenberg and Swendsen, 1988; Kumar et al., 1992, 1995] and MBAR [Shirts and Chodera, 2008] are used to produce reweighted analyses in the reference data and the validation data, respectively.</i>	44
3.6	<i>Heat capacity of the GNNQQNY-15 system with a 40 Å periodic simulation box. Far greater sampling than the original simulation leads to the sharp peak seen in the validation dataset.</i>	45
3.7	<i>Validation data for the ImpSTMD software, comparing properties of the BLN 48-mer to equivalent work in the reference paper [Kim et al., 2007]. The original authors distinguish between staircase and linear interpolation of their running statistical temperature estimates in fig. 3.7a and 3.7b, however this does not produce a visible difference. The current work uses linear interpolation.</i>	48
3.8	<i>Data snapshot from an STMD simulation of the 38-atom Lennard-Jones cluster system. Statistical temperature estimates are very far from convergence, as the simulation quickly became stuck in a local minimum configuration. This resulted in repeated erroneous revisions of the statistical temperature estimates adjacent to this bin, according to equation (2.23). The visits histogram is reset the first time a bin's statistical temperature estimate reaches the lower bound, and this is why all other bins are at zero visits.</i>	50

3.9	<i>Two STMD simulations of a single n16N peptide in a large periodic box. 100 bins span the potential energy range $[-80, 125]$, of which the sampled subsection is shown. The smoother dataset is not free of fluctuations and took 3×10^9 time-steps to reach; long enough to get a thorough sample of the system from REMD.</i>	51
4.1	<i>The four top-occurring structures for n16N represented in PLUM at 300.0K are displayed. Each structure is labelled by its rank, with the percentage of frames conforming to the structure parenthesised. The N-terminus is highlighted red and kept on the right for clarity. These top four structures cumulatively occupy 67.0% of all frames with highly ordered helical structure.</i>	55
4.2	<i>Ramachandran plots showing the exploration of (ϕ, ψ) phase space for a single unit of n16N at various thermostatted temperatures. Even at 325K, the α-helix peak is dominant and thin. Some exploration of unfolded states and left-handed α-helix occurs at all temperatures shown, but an extremely high temperature of 350K is required for the global maximum to be in an unfolded region. This makes it clear that the PLUM model is over-stabilising this form of secondary structure for the n16N molecule.</i>	56
4.3	<i>Behaviour of the PLUM model of n16N at 300K as a function of hydrogen bond interaction strength ϵ_{HB}, compared to an atomistic REMD simulation [Brown et al., 2014] with the CHARMM22* model [Piana et al., 2011; MacKerell et al., 1998] in TIPS3P water [Jorgensen et al., 1983]. The PLUM output is very sensitive to adjustments, and reaches peak similarity to the atomistic data with a decrease of about 5%. Each graph has \times symbols on the top axis showing the temperatures of simulations which were run.</i>	57
4.4	<i>The four top-occurring structures for n16N represented in PLUM* at 300.0K are shown. The backbone hydrogen bonding parameter has been reduced to 94.5% of its canonical value. Each structure is labelled by its rank, with the percentage of frames conforming to the respective structure parenthesised. The N-terminus is highlighted red and kept on the right for clarity.</i>	59

4.5	<i>Ramachandran plot for a single unit of n16N at 300.0K in the PLUM* model, where the backbone hydrogen bonding strength parameter ϵ_{HB} is set to 94.5% of its original value. This confirms that a higher degree of disorder now occurs, though no new peaks emerge.</i>	60
4.6	<i>The degree of manifestation of two different forms of secondary structure; α-helix and, broadly, “top left quadrant”, for each individual peptide bond along the chain. Proposed subdomains are demarcated by vertical dashed lines. The most striking resemblance is that each top left quadrant line is punctuated by two valleys centred on glycines, fluctuating about 0.6 otherwise. Each α line hits a minimum around SD2’s I residue, however, the disagreement in relative α-helicity between each subdomain is clear. For each pair of lines, several other minor peaks and troughs appear to line up well.</i>	60
4.7	<i>Secondary structure content of simulations of (a) n16N in the CHARMM22* model [Brown et al., 2014], (b) n16N in the PLUM* model, and (c) n16NN in the PLUM* model. Data represents occupancy of Ramachandran regions, without implying stable structure. The majority of segments match well in (a) and (b), but PLUM* has greater γ-structure and other structure, at the expense of PPII structure. n16NN has an increased propensity for α-structure, at the expense of most other structure forms. α_{left} may be under-represented in the PLUM* data, because of differences in the (ϕ, ψ) angles involved in the two models, and the authors’ choice of α_{left} region, provided in their fig. S1.</i>	61
4.8	<i>S1 Ramachandran plots from simulation, showing a deficiency in the PLUM model for this peptide. The ideal PPII peak, present in fig. 4.8a, is accessible but not favoured in the PLUM and PLUM* simulations of this peptide.</i>	63
4.9	<i>Ramachandran plot for a single unit of n16N in the PRIME20-like model. The temperatures in reduced units are: (top left) 0.11, (bottom left) 0.16 and (right) 0.135. Two peaks are notable; one indicating anti-parallel beta structure and another at $(-121, -28)$ which does not correspond to α-structure but appears to be a turn.</i>	66

4.10	<i>Occupancy of each of the four Ramachandran quadrants as a function of temperature. Crosses are shown on the top axis to indicate the temperatures at which data was collected. To enable comparison, atomistic data for n16N at $T = 300K$ is also shown [Brown et al., 2014]. This coarse measure of structure does not show signs that the peptide changes significantly as a function of temperature. Fig. 4.9 and fig. 4.11 show that the present level of bottom left quadrant structure is disproportionately not α-structure.</i>	66
4.11	<i>The prevalence of two different forms of secondary structure, α-helix and, broadly, “top left quadrant”, for each individual peptide bond in n16N simulated at $T^* = 0.135$. Proposed subdomains are demarcated by vertical dashed lines. As with the PLUM* n16N data (fig. 4.6) valleys exist in the top left quadrant lines around glycine. SD1 and SD2 agree rather well, but, also in common with PLUM*, a lower level of α-helical structure manifests in SD3 in the present model than in the atomistic data.</i>	67
4.12	<i>The four top-occurring structures for n16N represented in the PRIME20-like model at $T^* = 0.135$ are shown. Each structure is labelled by its rank, with the percentage of frames conforming to the respective structure parenthesised. For clarity, side-chains are not shown, the backbone is coloured yellow for residues involved in backbone-backbone hydrogen bonds, and the N-terminus is highlighted red and kept on the right.</i>	68
4.13	<i>Ramachandran plot of the n16NN peptide at $T^* = 0.135$ in the PRIME20-like model. A peak spanning a $\delta\psi$ of almost 100° is present in the β-sheet region while a much less populous peak occurs in the α-helix zone, unlike for the n16N peptide (fig. 4.9).</i>	70
4.14	<i>Top structure of the cluster analysis on the n16NN peptide in the PRIME20-like model. Side-chains are not shown, and additional visual cueing is provided by the x-coordinate dependent colour scheme. The N-terminus is shown in red in the bottom right. This cluster has a frame population of 23.1% and shows an irregular helix-like structure with a large diameter.</i>	71

5.1	<i>The four top-occurring structures for the n16N-2 system in PLUM* at 300.0K. Each structure is labelled by its rank, with the percentage population given in brackets. N-termini are highlighted in red. In all structures, the N-terminal half is central to the structure while the C-terminal half forms a tail. Differences in side-chain interactions must be responsible for this asymmetry.</i>	75
5.2	<i>The proportion of trajectory snapshots for which any given residue along the chain is involved in an interaction binding it to the other chain, in the n16N-2 and n16NN-2 systems. The n16N residue sequence is shown on the x-axis; the red residues are replaced in n16NN according to D→ N and E→ Q. Interactions are divided into side-chain and hydrogen bond types; an interaction for glycine is not always applicable, and in these cases no data-point is plotted. Note that the disparate forms of interaction in each model make a comparison of the average of each line meaningless.</i>	76
5.3	<i>Ramachandran plots of the n16N-2 system. Left shows a standard Ramachandran heat map. The greatest and widest peak is for β-strand structure, while a minuscule peak exists for α-helix structure, with a strong pathway between the two. Right shows a difference heat map, with the Ramachandran heat map of PLUM* n16N-1 (fig 4.5) normalised appropriately and subtracted from the heat map plotted on the left. In the scale shown, a value of -1.0 would imply 100% of hits being in a given bin in the n16N-1 simulation, and 0% in the n16N-2 simulation. Two absences are revealed in the locations for α and α_{left} structure, while β-structure is far more prominent. This implies that β-structures involving interpeptide interactions are more stable than intrinsically intrapeptide α-helices.</i>	77
5.4	<i>Ramachandran heat map for the n16N-2 system in the unaltered PLUM model. Unlike in the PLUM* model, n16N in PLUM does not shift away from α-helix structure in the dimer system.</i>	78
5.5	<i>The four top-occurring structures for the n16N-2 system in PRIME20-like model at $T^* = 0.135$. Each structure is labelled by its rank, and the percentage population is given in brackets. N-termini are highlighted in red. Interpeptide and intrapeptide hydrogen bonding are highlighted in black and yellow respectively.</i>	79

5.6	<i>Ramachandran plots of the n16N-2 system. Left shows a standard Ramachandran heat map. As with n16N-1 in PRIME20-like (fig. 4.9), there are two notable peaks, one indicating anti-parallel β-structure, and the other which has been characterised as a turn. Compared to n16N-1, the peaks have broadened greatly. Right shows a difference heat map, highlighting the difference between the n16N-1 and n16N-2 systems in (ϕ, ψ) angles adopted. While it appears that some systematic changes have occurred, these are much harder to distinguish from noise than the PLUM* case (fig. 5.3), and may simply be peak broadening. The differences also have about a fifth the magnitude of the PLUM* case.</i>	80
5.7	<i>Ramachandran heat map for the n16NN-2 system in the PLUM* model at 300 K. The global maximum exists in a sharp peak for α-helix structure, and a broader peak exists for β-structure.</i>	82
5.8	<i>Ramachandran heat map for the n16NN-2 system in the PLUM* model at 300 K. The largest island of allowed (ϕ, ψ) values is extremely flat, having no strongly preferred regions, though its maximum is in the location of anti-parallel β-structure, as with n16N-2.</i>	83
5.9	<i>The four top-occurring structures for the n16N-3 system in PLUM* at 300.0K. Each structure is labelled by its rank, with the percentage population given in brackets. N-termini are highlighted in yellow. At an RMSD cut-off of 0.8 nm, this analysis coarsely groups frames in which chains are similarly positioned, with little discrimination based on local intrapeptide secondary motifs.</i>	86
5.10	<i>Ramachandran plots of the n16N-3 system. Left shows a standard Ramachandran heat map. The system is strongly β-structure dominated. Right shows a difference heat map, comparing the (ϕ, ψ) coordinates visited with those of the n16N-2 system in PLUM*. The map is similar to the difference map of n16N-1 and n16N-2, shown in fig. 5.3, though the magnitude of the changes is far smaller.</i>	87

5.11	<i>The proportion of trajectory snapshots for which any given residue along the chain is involved in an interaction binding it to an other chain, in the n16N-3 and n16NN-3 systems. The n16N residue sequence is shown on the x-axis; the red residues are replaced in n16NN according to $D \rightarrow N$ and $E \rightarrow Q$. Interactions are divided into side-chain and hydrogen bond types; an interaction for glycine is not always applicable, and in these cases no data-point is plotted. Note that the disparate forms of interaction in each model make a comparison of the average of each line meaningless.</i>	88
5.12	<i>The four top-occurring structures for the n16N-3 system in the PRIME20-like model at $T^* = 0.135$. Each structure is labelled by its rank, with the percentage population given in brackets. N-termini are highlighted in yellow.</i>	89
5.13	<i>Ramachandran plots of the n16N-3 system in PRIME20-like. Left shows a standard Ramachandran heat map. The global maximum is now the turn structure in the bottom left of the accessible region. Two significant, broad peaks also exist at low and high values of ϕ in the β-structure domain. Right shows a difference heat map, comparing the (ϕ, ψ) coordinates visited with those of the n16N-2 system in PRIME20-like. Within the top left quadrant, the peak for lower values of ϕ has weakened, while the peak for higher values has increased, as it did between n16N-1 and n16N-2. The trends concerning the bottom left quadrant have reversed, however.</i>	90
5.14	<i>The two top-occurring structures for the n16N-6 system in PLUM* at 300.0 K. Each structure is labelled by its rank, with the percentage population given in brackets. N-termini are highlighted in lime.</i>	93
5.15	<i>Ramachandran plots of the n16N-6 system. Left shows a standard Ramachandran heat map. The plot is dominated by the β-structure peak, which is twice the magnitude of any other peak. A wide range of other angle coordinates are accessible. Right shows a difference heat map, revealing the difference between the normalised n16N-6 and n16N-3 systems. Continuing the trend from n16N-1 to n16N-2 to n16N-3, most accessible regions and particularly the α-helix region have drained, while the β-structure peak has risen. However, new peaks can be seen growing at coordinates $(-50^\circ, -100^\circ)$ and $(40^\circ, 110^\circ)$. These dihedral angle pairs occur in SD1 and SD2 at the start and end of β-strands.</i>	94

5.16	<i>The proportion of trajectory snapshots for which any given residue along the chain is involved in an interaction binding it to an other chain, in the n16N-6 and n16NN-6 systems. The n16N residue sequence is shown on the x-axis; the red residues are replaced in n16NN according to $D \rightarrow N$ and $E \rightarrow Q$. Interactions are divided into side-chain and hydrogen bond types; an interaction for glycine is not always applicable, and in these cases no data-point is plotted. Note that the disparate forms of interaction in each model make a comparison of the average of each line meaningless.</i>	95
5.17	<i>The two top-occurring structures for the n16N-6 system in PRIME20-like at $T^* = 0.135$. Each structure is labelled by its rank, with the percentage population given in brackets. N-termini are highlighted in lime.</i>	96
5.18	<i>Ramachandran plots of the n16N-6 system in PRIME20-like. Left shows a standard Ramachandran heat map. Within the island of sterically accessible dihedral angles, no coordinates are strongly favoured or disfavoured. Right shows a difference heat map, comparing the (ϕ, ψ) coordinates visited with those of the n16N-3 system in PRIME20-like. While the changes seen here are reversed compared to the changes between the dimer and trimer n16N systems, the magnitude of changes observed remains very low.</i>	97
5.19	<i>A heat map illustrating the frequency of interpeptide residue-residue contacts in the PRIME20-like model's n16N-6 system. The dataset plotted includes all non-bonded interactions except for backbone hydrogen bonding. A value of 1.0 would indicate two residues interacting in as many pairs as possible (15 pairs for a 6-chain system) in every frame. On this scale, a value of 0.16 implies an average of 2.4 corresponding residue pairs are in contact.</i>	98
5.20	<i>The topmost populated geometric cluster of the n16NN-6 system in PLUM*, with a population of 10.1%. Four chains form a parallel β-sheet, while the other two are tightly packed in a hybrid β-strand/β-hairpin conformation which involves both inter- and intra-peptide hydrogen bonding. SD3 of the red chain joins the β-sheet, which would be very unusual for n16N.</i>	99

Acknowledgments

I would like to begin by thanking my supervisors, Mike Allen and David Quigley, for their tireless support and direction, in matters of both programming and science. I would like to thank Marcus Bannerman for his guidance and substantial contributions in getting the PRIME20-like discontinuous protein model implemented. I gratefully acknowledge the help of Tiff Walsh and Aaron Brown for sharing data and ideas as we considered the n16N peptide together, and Matt Bano who suggested studying the n16NN peptide.

Thanks to my office mates, Štěpán Růžička, Sam Brown, Mike Ambler, and Poppy Asman for academic assistance, and for ensuring the experience was lively. I would like to express my feeling of indebtedness to the wider Warwick theory group, who have always maintained a strong atmosphere of camaraderie and mutual advocacy. The MIB consortium, which my project brought me in to, showed me consistent friendliness and support, which I am grateful for and hope to have reciprocated. I would particularly like to recognise the aid and friendship of Jasmine Desmond from that group. Lastly, I would like to extend my gratitude beyond those I know directly through the scientific community, to friends who have accompanied me on my journey these past years, and to my family, whose love I have always been able to count on.

Declarations

The results and computer program extensions presented as new in this thesis are my original work. They were produced between October 2011 and May 2015, during the term of my Ph.D. course, under supervision from Professor M. P. Allen and Dr. D. Quigley. Where work, information and tools have come from outside sources, every effort has been made to make this known. No part of this work has been previously submitted to the University of Warwick, nor to any other academic institution, for the purpose of obtaining a higher degree.

Abstract

Intrinsically disordered proteins (IDPs) are functional proteins which lack a unique and stable tertiary structure. IDPs such as n16N are involved in biomineralisation, the process by which organisms produce mineral materials, such as shells. Here, the role that accelerated simulation can play in the study of IDPs is examined and furthered. The coarse-grained models PLUM and PRIME20 are implemented and refined based on existing single-chain n16N simulations. In conjunction with the replica exchange molecular dynamics technique, the models are used to simulate systems of 1, 2, 3 and 6 chains of n16N, and a mutant form n16NN. The modified PLUM model is in striking agreement with existing hypotheses regarding the structure of n16N, when simulations are run in multiplicity. The PRIME20 model has difficulty producing plausible backbone structure in every system size, though it does fulfil some expectations regarding residue interaction specificity. New hypotheses are offered on bulk n16N and n16NN aggregation based on the presented data. Future directions for development of accelerated simulation techniques for IDPs are suggested.

Chapter 1

Introduction

This thesis looks at the use of molecular dynamics with coarse-grained models to simulate intrinsically disordered proteins involved in biomineralisation. The goal is to aid the effort to understand biomineralisation proteins and to advance the role of simulation in the study of disordered proteins. This introduction serves to amalgamate the background knowledge which makes the motivation for the project clear.

Building up from a background of protein order and disorder in sec. 1.1 and sec. 1.2 respectively, an introduction and current work in relevant areas of biomineralisation research are given in sec. 1.5, where the central protein of this project, n16N, is also introduced. Sec. 1.3 surveys the experimental research methods used to understand protein disorder and proteins in biomineralisation systems, while sec. 1.4 presents the theory behind applying computer simulation to the problem of protein function. Chapter 2 will subsequently provide an in-depth look at how it is hoped molecular dynamics can be accelerated in its sampling of complex protein systems, through coarse-graining and other methods, concluding the introductory content of the thesis.

1.1 Protein structure

Proteins are biological polymers consisting of combinations of 20 different amino acid residues. Depending on the amino acid sequence, proteins have the capacity to take on a vast number of three-dimensional structures and functions.

Protein structure is classified in four levels. The primary structure of a protein is the linear sequence of amino acids from which it is made, and can be given as a simple sequence of letters, each denoting an amino acid. An amino acid

chain will have a free amine group on one end and a free carboxyl group on the other; these are the N- and C- termini respectively, and the primary structure is specified from the N-terminus to the C-terminus.

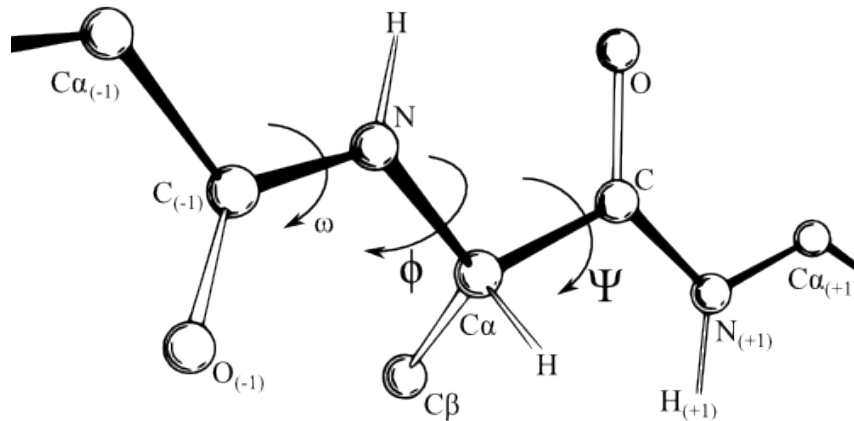


Figure 1.1: Wikimedia user Dcrjrsr, available under the Creative Commons Attribution 3.0 Unported license.

The backbone atoms of a central amino acid residue is shown, with two neighbours labelled (-1) and (+1) partially shown. Each residue except glycine also has a side-chain, denoted $C\beta$. The dihedral angles ϕ , ψ and ω which determine the secondary structure are labelled.

Regardless of the primary structure, the chain of amino acids has a repeating backbone as shown in fig. 1.1. The backbone's flexibility is due to its dihedral angles ϕ and ψ , while ω is inflexible. Steric considerations and backbone interactions pre-sculpt the free energy landscape, creating limited favoured possibilities for ϕ and ψ [Hoang et al., 2004; Ho et al., 2003]. The specific properties of a residue are due to the chemistry of its unique side-chain group, bonded to the central carbon. The pre-sculpting and side-chain interactions together typically yield one or more local native conformations for the chain, and these motifs are known as secondary structure.

Motifs involve a regularly repeating structure, through values of ϕ and ψ which repeat for some number of residues along the chain. The most common motifs are β -sheets and α -helices. β -sheets are formed from elongated strands of the chain running in parallel or anti-parallel, laterally joined by hydrogen bonds. The α -helix is a coiled conformation in which each NH backbone group is hydrogen bonded to the CO group of the amino acid four residues prior (denoted $i + 4 \rightarrow i$). Different amino acid residues have varying propensities towards the possible secondary structural

forms, and table 1.1 shows this for the two most common cases.

The allowed values of ϕ and ψ can be visualised on a 2D plot known as a Ramachandran plot, and fig. 1.2 illustrates this. Accumulating sufficient (ϕ, ψ) coordinates from a meaningful dataset such as the residues in a portion of a protein can lead to a probability density function, which can be plotted as a heat map Ramachandran plot, to better characterise the secondary structure.

The exclusionary zones seen on a Ramachandran plot can be classified by the steric clashes which lead to them. Ramachandran laid out his suggested classification [Ramachandran et al., 1963], but fig. 1.3 presents an updated view.

Tertiary structure describes the geometric shape of the protein; how the secondary structural motifs and disordered regions fold together. This is largely determined by the intra-protein interactions between side-chains. Finally, quaternary structure describes the geometry of multiple folded chains together.

Beyond the covalent bonds, the most important interactions are backbone-to-backbone hydrogen bonding for determination of secondary structure, and side-chain and solvent hydrogen bonding, hydrophobic side-chain interactions, ionic side-chain interactions, and disulfide bridges between cysteine groups for determination of tertiary and quaternary structure.

1.1.1 Dependence on ambient conditions

Solvent, ionic strength, other molecules, interfaces, pressure and temperature are some of the ambient factors which impact the structure a protein adopts.

Favoured protein conformations tend to minimise disruption of the water matrix [Fernández, 2013]. Coulombic interactions are screened by the water by a factor of about 80 [Berg et al., 2002, page 11], while free ions strongly screen Coulomb interactions. Water's ability to hydrogen bond creates competition between solute-solute and solute-solvent hydrogen bonding. Conversely, hydrophobic residues get buried inside the protein core to avoid water.

Proteins are extremely temperature-sensitive and *denature* at excessive temperatures. Denaturation is the permanent loss of the native state, often accompanied by aggregation and a loss of solubility. This typically occurs within a few tens of degrees Celsius of the organism's temperature of adaptation (many examples in [Somero, 1995]); organisms containing proteins that can retain structure at temperatures of 90-100°C are those which have a need to survive such conditions [Somero, 1995]. The evolutionary cause of this may be the requirement of rapid, reversible changes to protein conformation for most proteins [Creighton, 1984, page 507]. Extremely minor protein sequence changes increasing internal hy-

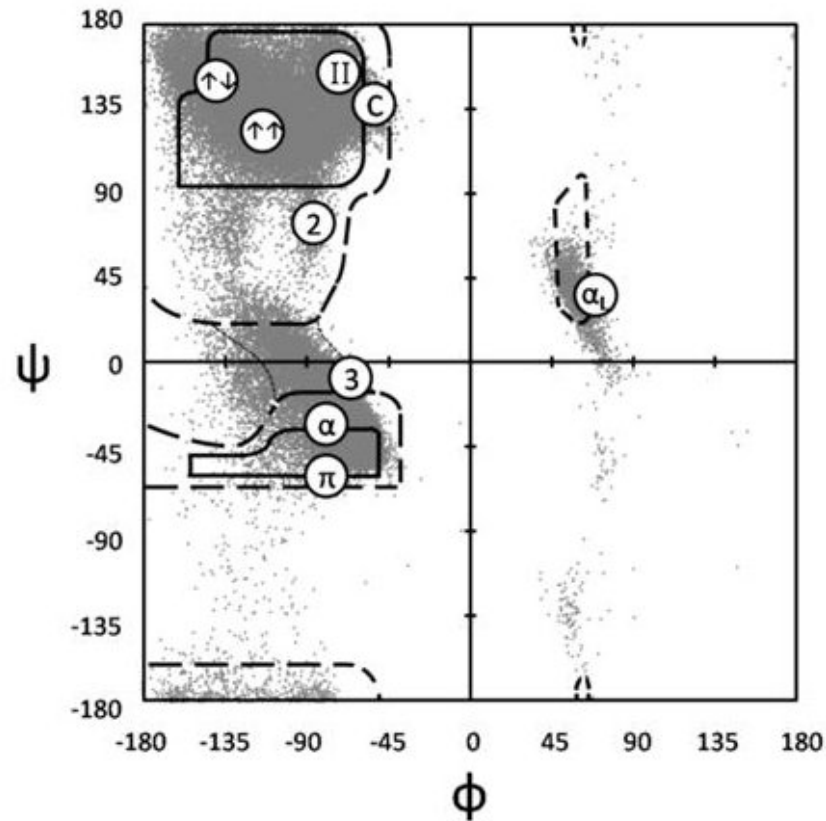


Figure 1.2: [Hollingsworth and Karplus, 2010] adapted from [Ramachandran and Sasisekharan, 1968].

A Ramachandran plot showing both example data and outlines of commonly accepted regions. The data is comprised of 63,149 residues from crystal structures and each point indicates the (ϕ, ψ) values representing the secondary structure of a single residue.

The outlines are divided into core allowed regions (solid lines) and allowed regions (dashed lines). Several forms of secondary structure have their location indicated; these are α -helix (α), 3_{10} -helix (3), π -helix (π), left-handed α -helix (α_1), 2.27 ribbon (2), polyproline-II (II), collagen (C), parallel β -sheet ($\uparrow\uparrow$) and anti-parallel β -sheet ($\uparrow\downarrow$).

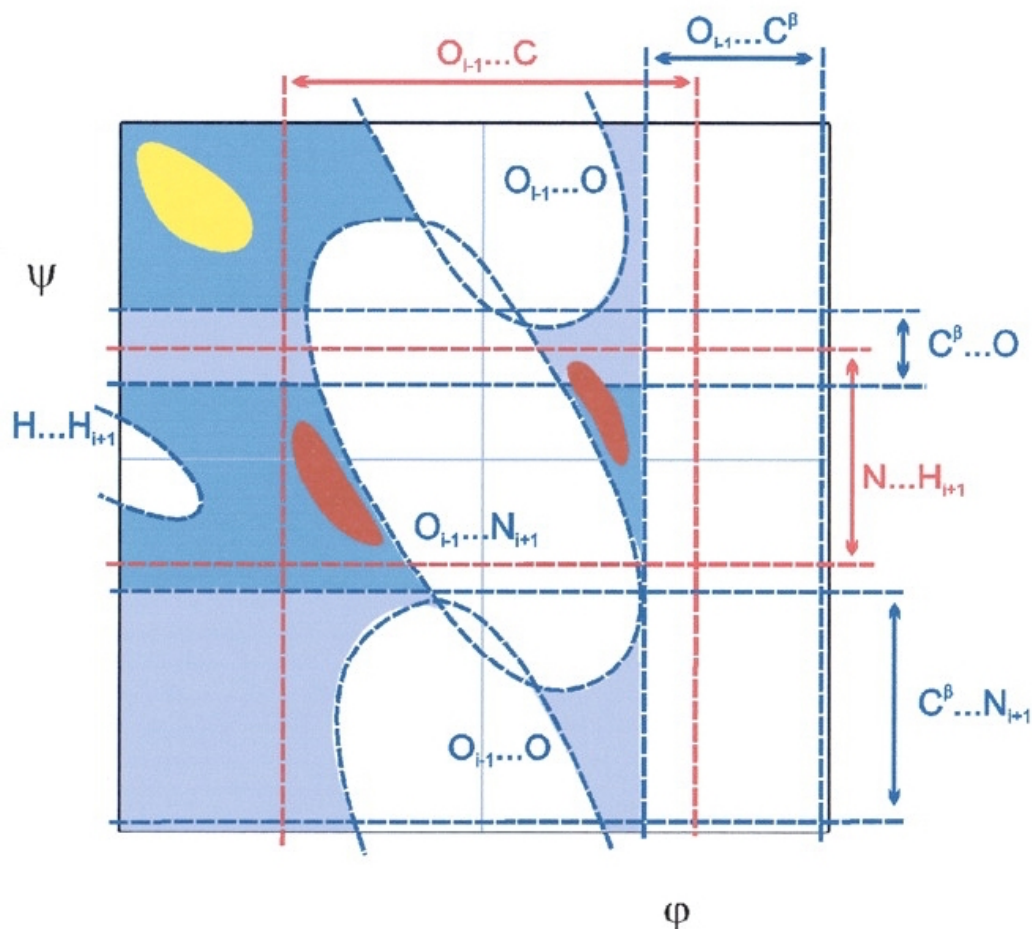


Figure 1.3: A steric map, in which steric clashes leading to forbidden regions are shown in (ϕ, ψ) phase space and labelled according to the clashing pair of atoms [Ho et al., 2003]. Dashed blue lines and blue labels denote clashes. Favourable dipole-dipole backbone interactions are also plotted, with red dashed lines and red labels. Areas of the map are white if forbidden, and otherwise coloured according to the legend below. The atom labels match the illustration in fig. 1.1.

Boundaries

- - - Attractive dipole-dipole interactions
- - - Steric clashes

Accessible regions

- Sterically permitted
- Single steric clash (outlier region)
- Left- and right-handed α -helix regions
- β -strand region

drophobicity and stabilising specific secondary structural motifs result in significant temperature fortification [Imanaka et al., 1986; Matthews, 1993].

1.2 Protein disorder

Normal proteins are thought to have strongly funneled energy landscapes, leading them to folded native states with an energetically favourable fixed structure on every level [Wolynes et al., 1995; Baker, 2000]. The classical structure-function paradigm of proteins emerged from the proposition that protein denaturation is purely a conformational change [Wu, 1931], and states that 3D protein structure determines its function, therefore, all functional proteins require a fixed native state. The cataloguing of thousands of functional native protein structures during the rest of the 20th century [Berman et al., 2000] concreted the notion of 3D structure being a prerequisite of function.

Evidence of configurational adaptability [Karush, 1950; Bennett and Steitz, 1978] and functional disordered regions (numerous examples cited in [Plaxco and Groβ, 1997]) also crept in during the second half of the 20th century, though it took until the turn of the millennium for papers to appear formally arguing for function in natively disordered proteins [Wright and Dyson, 1999; Uversky et al., 2000; Dunker et al., 2001; Tompa, 2002], and casting doubt on the universality of the structure-function paradigm.

IDPs (intrinsically disordered proteins) are defined by their inability to fold into a unique and stable tertiary structure, and this term is preferable to the early term IUPs (intrinsically unstructured proteins) which falsely suggests a complete lack of structure [Dunker et al., 2013].

1.2.1 Determinants of protein disorder

The disorder-promoting amino acid residues are roughly agreed-upon as being A, G, P, R, Q, S, E and K [Williams et al., 2001; Campen et al., 2008; Dunker et al., 2001]. Table 1.1 includes a ranking of amino acid disorder propensity. Amino acids over-represented in IDPs are polar, charged and not hydrophobic [Rani et al., 2014].

The “sequence complexity” of proteins [Romero et al., 2001] is defined using Shannon’s information entropy [Shannon, 1948] as

$$S = - \sum_{i=1}^N f_i (\log_2 f_i) \quad (1.1)$$

in which N is the number of types of amino acids, i.e. 20, and f_i is the fraction of

Amino acid	Abbreviations	α -helix	β -sheet	Disorder
Alanine	Ala, A	1.29	0.90	0.060
Arginine	Arg, R	0.96	0.99	0.180
Asparagine	Asn, N	0.90	0.76	0.007
Aspartic Acid	Asp, D	1.04	0.72	0.192
Cysteine	Cys, C	1.11	0.74	0.020
Glutamine	Gln, Q	1.27	0.80	0.318
Glutamic Acid	Glu, E	1.44	0.75	0.736
Glycine	Gly, G	0.56	0.56	0.166
Histidine	His, H	1.22	1.08	0.303
Isoleucine	Ile, I	0.97	1.45	-0.486
Leucine	Leu, L	1.30	1.02	-0.326
Lysine	Lys, K	1.23	0.77	0.586
Methionine	Met, M	1.47	0.97	-0.397
Phenylalanine	Phe, F	1.07	1.32	-0.697
Proline	Pro, P	0.52	0.64	0.987
Serine	Ser, S	0.82	0.95	0.341
Threonine	Thr, T	0.82	1.21	0.059
Tryptophan	Trp, W	0.99	1.14	-0.884
Tyrosine	Tyr, Y	0.72	1.25	-0.510
Valine	Val, V	0.91	1.49	-0.121

Table 1.1: *Relative frequencies of residues in common secondary structure motifs [Creighton, 1992] [Berg et al., 2002, page 67] and disorder propensity according to the TOP-IDP scale [Campen et al., 2008], in which higher values are more disordered. This is one of many scales on which residue disorder propensity has been ranked [Galzitskaya et al., 2006; Oldfield et al., 2005; Vihinen et al., 1994; Garbuzynskiy et al., 2004]. The highest values are highlighted.*

amino acid type i which appears in a window of the chain of length L . There is strong correlation between protein disorder and *low* sequence complexity [Romero et al., 2001; Rani et al., 2014].

Bioinformatic disorder predicting algorithms have been developed [He et al., 2009], with meta-predictors recently approaching 90% accuracy [Monastyrskyy et al., 2014]. Such tools allow us to project that 15-45% of proteins in eukaryotes contain disordered regions at least 30 residues in length [Tompa, 2012].

1.2.2 Functions of IDPs

Functions of IDPs have been grouped into four broad categories [Dunker et al., 2002]: entropic chains, molecular recognition, protein modification, and molecular assembly.

Entropic chains have functions dependent directly on their disorder: entropic springs exhibit an entropic force according to Hooke’s law, the equilibrium position corresponding to a maximum of entropy; entropic spacers flexibly link functional subunits of large proteins; and entropic clocks are commonly found in ion channels and take some characteristic time of random exploration to find and bind to a target, opening or closing the channel. IDPs are perfect candidates for *molecular recognition* of multiple heterogeneous targets with low affinity, useful for signalling purposes. Intrinsic disorder has been shown to play an important role in some post-translational *protein modifications* [Gao and Xu, 2012], both by protease cleavage or chemical additions.

The last category is *molecular assembly*, which is crucially relevant to extracellular biomineralisation, described in Sec. 1.5. Complex frameworks composed of protein subunits may benefit greatly from disorder in their subunits for multiple reasons. Complexes may often be assembled in multiple stages, making great use of conformational flexibility of the subunits to overcome steric barriers to formation. Unfolding and refolding requires relatively weak binding interactions, yet stable attachment will be guaranteed in a macromolecule by the large number of these interactions. An IDP may exhibit selectivity of the chemical environment in which it folds, and the environment can even affect the properties of the complex framework arising from a given IDP [Namba, 2001].

1.3 Experimental techniques

Several methods have been used to determine and characterise intrinsic disorder and study biomineralisation systems.

1.3.1 Intrinsic disorder

X-ray crystallography is a highly favoured technique for structural characterisation of crystalline matter, and can be used to study proteins. However, disorder leads to missing residues in structures determined by x-ray crystallography, providing a hint at disorder but no means to characterise it in any detail [Dunker et al., 2001]. NMR has played a key role in IDP structure and dynamics studies [Showalter, 2007].

Several other techniques have played smaller roles. Small-angle X-ray scattering has the capacity for “automated and rapid characterisation of protein solutions in terms of low-resolution models, quaternary structure and oligomeric composition” [Mertens and Svergun, 2010] and has been advocated [Sibille and Bernado, 2012]

and used [Sterckx et al., 2014; Wells et al., 2008; Zhang et al., 2013] as a complement to NMR for IDPs.

Other techniques of interest include circular dichroism (CD), fluorescence spectroscopy and hydrodynamic techniques.

NMR spectroscopy

Atomic nuclei with odd numbers of protons and/or neutrons have a nuclear spin (I) which causes them to interact with applied magnetic fields. When a magnetic field of strength B_0 is applied, the most salient (spin- $\frac{1}{2}$) cases of ^1H and ^{13}C will orient themselves in one of two ways; aligned with or opposed to the field. The difference in energy between these two states is given by

$$\Delta E = \gamma \hbar B_0 \quad (1.2)$$

where γ is the magnetogyric ratio, a property of an isotope. Nuclei in the magnetic field will naturally precess with an angular frequency known as the Larmor frequency given by

$$\omega = -\gamma B_0 . \quad (1.3)$$

A radio wave with a matching frequency will be partially absorbed and excite some of the nuclei to the high-energy state. A short pulse of some μs covering a band of frequencies will excite nuclei with a matching band of excitation frequencies. The nuclei will immediately begin to relax back to equilibrium excitation levels given by the Boltzmann distribution:

$$\frac{N_\beta}{N_\alpha} = \exp \frac{-\Delta E}{kT} \quad (1.4)$$

where N_β is the number of nuclei in the high energy state, and N_α the low energy state. During relaxation, the non-equilibrium magnetisation precessing about the external magnetic field can induce an electromotive force known as the free induction decay (FID). The FID is received as a signal, and Fourier transformed to provide information on the frequencies which were excited.

All nuclei of the same isotope will have similar resonance frequencies, but variations in the electron density of the environment of each nucleus will lead to small deviations known as the *chemical shift*. Thus, NMR can distinguish between ^{13}C atoms occurring in different chemical environments, and deduce molecular structures [Williams and Fleming, 1995].

NMR has been used to study the protein folding process and unfolded proteins [Dyson and Wright, 2004, 2002; Brockwell et al., 2000] and development of NMR techniques for IDPs is a fertile field [Bertini et al., 2011; Jensen et al., 2013; Ota et al., 2013; Konrat, 2014; Felli and Pierattelli, 2014]. In particular, there is a need for software to generating and validating IDP conformation ensembles from NMR data to catch up [Showalter, 2007], perhaps involving a dual NMR-simulation approach [Ball et al., 2013].

No NMR or SAXS studies of n16N in solution have been found in the literature, but such studies could provide interesting insights into n16N's ensemble of structures and the macromolecular complex it forms.

1.3.2 Biomineralisation systems

In vitro recreation of aspects or the whole of an organic biomineralisation environment, using experimental tools to investigate what occurs, is a very common research strategy for n16N [Seto et al., 2014; Amos et al., 2011; Keene et al., 2010a,b; Metzler et al., 2010; Delak et al., 2007] and in general. The findings on n16N discussed in section 1.6 come from this category of study.

In these studies, n16N is allowed to aggregate in solution or on a substrate. Studies may proceed by observing n16N aggregation, studying aggregates, introducing Ca^{2+} and altering its concentration, observing the response to pH changes, observing calcium carbonate crystal growth, introducing silk fibroin gel, or in other ways. A wide range of experimental techniques can be used to inspect the system. Table 1.2 summarises the methods used by studies conducted so far.

Citation	Salient assay properties	Techniques used to analyse...	
		Organic complexes	(Bio)mineral products
[Kim et al., 2004a]	1: Overgrowth. Geological calcite 2: Kevlar fibres	CD	SEM, XRD
[Kim et al., 2006]	Calcite crystals, growth solution	CD	AFM
[Delak et al., 2007]	Kevlar fibres	CD	SEM, EDX
[Metzler et al., 2008]	Kevlar fibres	-	XANES
[Keene et al., 2010a]	β -chitin substrate	FM	SEM, XRD, Raman
[Metzler et al., 2010]	Kevlar fibres	X-PEEM	X-PEEM
[Keene et al., 2010b]	β -chitin substrate, silk fibroin	FM	SEM, Raman, EDX
[Amos et al., 2011]	1: Excess CO ₂ to manipulate pH 2: Pre-formed n16N complexes	CD, SEM, AFM, DLS, TEM	XRD, SEM
[Seto et al., 2014]	Titration of CaCl ₂ into buffer solution	TEM	SEM, TEM, FTIR, XRD

Table 1.2: Summary of experimental research on n16N peptide, with a focus on experimental techniques used. All assays contain n16N in water solution with calcium and carbonate ions for crystal growth.

AFM - Atomic force microscopy
 CD - Circular dichroism spectroscopy
 DLS - Dynamic light scattering
 EDX - Energy dispersive X-ray spectroscopy
 FM - Fluorescence microscopy
 FTIR - Fourier-transformed infrared spectroscopy
 Raman - Raman spectroscopy
 SEM - Scanning electron microscopy
 TEM - Tunnelling electron microscopy
 XANES - X-ray absorption near edge spectromicroscopy
 XRD - X-ray diffraction
 X-PEEM - X-ray photoelectron emission spectromicroscopy

1.4 Molecular dynamics for proteins

While these experiments provide a great deal of insight into n16N’s abilities, it is very difficult to learn anything about what is happening at the atomistic scale, or even at the level of individual peptides, from real-world experiments. Simulations create a bridge between experiment and theory, in which hypotheses can be tested, altered, disproved, or corroborated, and a fuller picture of the behaviour of a biomineralisation system can be reached.

1.4.1 Approaches to molecular simulation

Molecular simulation comes in two key methods; these are Monte Carlo and molecular dynamics. The popular Metropolis algorithm for Monte Carlo [Metropolis et al., 1953] is a method of stepwise exploration of a system’s configurational space without regarding the dynamics of the system. Instead, a move set is defined which describes the ways the system’s configuration in the current step can be altered to produce the configuration in the next step. Moves are accepted with a probability given by

$$\mathcal{P} = \min(1, \exp(-\beta\delta\mathcal{V}_{nm})) \quad (1.5)$$

in which $\beta = 1/k_B T$ and $\delta\mathcal{V}_{nm} = \mathcal{V}(\mathbf{r}_n) - \mathcal{V}(\mathbf{r}_m)$ is the difference in configurational energy between the current configuration, \mathbf{r}_m and the next, \mathbf{r}_n . When a move is rejected, the current configuration is carried forward to the next step unaltered. It can be shown [Tildesley, 1993] that this generates states with the canonical probability distribution. In such a scheme, the thermodynamic average of a configurational property $\mathcal{A}(\mathbf{r})$ can be estimated as the straightforward average over the sampled states

$$\langle \mathcal{A} \rangle_{NVT} = \frac{\sum_{states} \mathcal{A}(\mathbf{r})}{N_{states}} \quad (1.6)$$

where \mathbf{r} is the set of $3N$ positions describing the system [Tildesley, 1993]. Thus, the statics of a molecular system can be characterised through the Monte Carlo method.

Molecular dynamics is an alternative to Monte Carlo that uses a model system which could be identical to one used in Monte Carlo, and advances the system’s positions and momenta in discretised time-steps δt through Newton’s equations of motion. Unlike in Monte Carlo, this requires calculation of the force \mathbf{f}_i on particle i of mass

m_i due to the potential \mathcal{V} :

$$\mathbf{f}_i = m_i \ddot{\mathbf{r}}_i = -\nabla \mathcal{V}(\mathbf{r}_i). \quad (1.7)$$

Several algorithms exist to increment from the positions of the current time-step $\mathbf{r}(t)$ to the new positions $\mathbf{r}(t + \delta t)$, the dominant one being the *velocity Verlet* algorithm [Swope et al., 1982]. This is a modification of the older *Verlet* algorithm [Verlet, 1967] derived from Taylor expanding $\mathbf{r}(t)$. In this formulation, positions and velocities are advanced together as follows

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2} \delta t^2 \mathbf{a}(t), \quad (1.8a)$$

$$\mathbf{v}(t + \delta t) = \mathbf{v}(t) + \frac{1}{2} \delta t (\mathbf{a}(t) + \mathbf{a}(t + \delta t)). \quad (1.8b)$$

with velocity $\mathbf{v}(t)$ and acceleration $\mathbf{a}(t)$. $\mathbf{a}(t + \delta t)$ must be evaluated between these two expressions.

1.4.2 Statistical ensembles

The Newtonian equations of motion conserve the total energy E of the system. In the simplest case, the simulation will also be time-independent in N , the system's particle count, and V , the system's volume. The set of microstates that are sampled in a system where this set of variables are independent is designated the *constant-NVE* or *microcanonical* ensemble. The microcanonical ensemble is defined by its probability density [Allen and Tildesley, 1987]

$$\rho_{NVE}(\mathbf{r}, \mathbf{p}) \propto \delta(\mathcal{H}(\mathbf{r}, \mathbf{p}) - E) \quad (1.9)$$

in which the Dirac delta function $\delta(x)$ is equal to 0 for $x \neq 0$, with an integral of 1 over the real number line. E is total energy which defines the ensemble, while the Hamiltonian $\mathcal{H}(\mathbf{r}, \mathbf{p})$ provides the total energy of microstate \mathbf{r} .

However, ensembles with varying E and specified thermodynamic (average) temperature T more closely resemble experimental and biological systems [Wang and McCammon, 2012, page 18]. The constant-*NVT* ensemble is known as the canonical ensemble and is defined by its probability density [Allen and Tildesley, 1987]

$$\rho_{NVT}(\mathbf{r}, \mathbf{p}) \propto \exp\left(\frac{-\mathcal{H}(\mathbf{r}, \mathbf{p})}{k_B T}\right). \quad (1.10)$$

We think of such a system as being composed of the experimental system

coupled to a large heat bath at the specified temperature T . For our outputted statistics to be of the canonical ensemble instead of the constant- NVE ensemble, some alteration of the equations of motion will be necessary; a temperature-preserving modification will be called a *thermostat*. Multiple thermostats which reproduce canonical statistics are available [Hünenberger, 2005] and two popular methods are *stochastic dynamics* and *stochastic coupling*.

The stochastic dynamics algorithm relies upon addition of terms from the Langevin equation to the force calculation in equation (1.7)

$$\mathbf{f}_i = -\nabla\mathcal{V}(\mathbf{r}_i) - m_i\gamma_i\mathbf{v}_i(t) + \eta_i(t). \quad (1.11)$$

$\eta_i(t)$ is a stochastic force uncorrelated with the velocities $\mathbf{v}_i(t')$ and systematic force $-\nabla\mathcal{V}(\mathbf{r}_i)$ at previous times $t' < t$, obeying

- $\langle\eta_i(t)\rangle = 0$, and
- $\langle\eta_{ai}(t)\eta_{bj}(t')\rangle = 2m_i\gamma_ik_B T\delta_{ij}\delta_{ab}\delta(t-t')$

in which subscript a and b index Cartesian axes. γ_i are damping constants, but the damping term can also be interpreted as a frictional term by setting $\zeta_i = \gamma_im_i$, in which ζ_i would be considered friction constants. However, in the absence of an external force, the velocity autocorrelation function is proportional to $\exp(-\gamma_it)$.

It can be shown that a trajectory generated by integrating the Langevin equation of motion produces a trajectory sampling from the canonical distribution [Hünenberger, 2005].

The stochastic coupling method is also known eponymously as the Andersen thermostat [Andersen, 1980]. Atoms are selected to have their velocity reassigned from the Maxwell-Boltzmann distribution corresponding to the selected temperature T , as if collisions were occurring with a heat bath at temperature T [Frenkel and Smit, 2002, page 142]. The selection process utilises a coupling strength parameter in the form of frequency of collisions, ν . Selections are carried out such that the probability of a time interval τ between two successive collisions is

$$P(\tau; \nu) = \nu \exp(-\nu\tau). \quad (1.12)$$

It can be shown [Andersen, 1980] that this scheme leads to a canonical distribution of microstates.

In the canonical ensemble, the constant-volume specific heat capacity is given

by the fluctuations in total energy E according to

$$\langle \delta E^2 \rangle_{NVT} = k_B T^2 C_V, \quad (1.13)$$

$C_V = \left(\frac{\delta E}{\delta T} \right)_V$ being the specific heat capacity [Allen and Tildesley, 1987]. The form of the heat capacity curve reveals the temperature of phase transitions in the system.

1.5 Biomineralisation

Biomineralisation is the term given to the formation of minerals in controlled environments by living organisms. Organisms display an ability to form nanostructured hard tissues through processes which are not well understood, but have a vast range of potential applications. A biomineral, the product of biomineralisation, is composed of mineral matter and organic matter. Until the early 1980s, the term *calcification* was used rather than *biomineralisation*, reflecting the fact that calcium carbonate is the most common mineral type.

Biomineralisation can be divided into *biologically induced* and *biologically controlled* mineralisation [Lowenstam, 1981; Mann, 1983]. In biologically induced mineralisation, the organism has little control over the type of mineral formation, although the organism exerts control over the chemical environment and nucleation event. Biominerals formed in this way have the hallmarks of inorganic minerals, such as poorly defined external morphology and heterogeneity of water content, composition and structure [Weiner and Dove, 2003].

Biologically controlled mineralisation denotes far finer control over the mineralisation process and the properties of the created biomineral. The mineralisation can occur intracellularly, involving nucleation in the cell, usually followed by excretion; intercellularly, in which the mineralisation site is isolated between cells; or extracellularly, in which a macromolecular complex is produced outside of cells and regulates the formation of the biomineral [Weiner and Dove, 2003].

Extracellular biomineralisation is a very common biomineralisation approach, used amongst many other examples to form shells of molluscs, cephalopod statoliths [Bettencourt and Guerra, 2000] and bones and teeth [He et al., 2003]. The macromolecule complex is comprised of proteins, polysaccharides or glycoproteins [Lowenstam and Weiner, 1989], with high levels of disorder in their native states; indeed, biomineralisation proteins have been called the most disordered functional class in the protein world [Kalmar et al., 2012]. Determining the functions of the matrix proteins is the bottleneck in understanding extracellular biomineralisation [Weiner and Dove, 2003].

The nacre layer of mollusc shells is a well-studied biomineral [Belcher et al., 1996; Thompson et al., 2000; Levi et al., 1998; Samata et al., 1999] made of calcium carbonate, through extracellular mineralisation. Fig. 1.4 shows a model of the macromolecular framework in which nacre is created. The involved components include *chitin*, an extremely abundant natural polymer which exists with a rigid crystalline structure. The polymorphs α and β are held together by intermolecular hydrogen bonding of chains in parallel or anti-parallel formations respectively [Kim et al., 1996].

Calcium carbonate is a polymorphic substance, able to exist in multiple crystal structures, the most stable three of which, in descending order, are calcite, aragonite and vaterite. Nacre exists as aragonite [Lowenstam and Weiner, 1989], indicating that the macromolecular environment has the capability of polymorph selectivity.

Indeed, the polymorph stabilisation achieved by the macromolecular components has been reproduced *in vitro*. A chitin-silk assembly similar to fig. 1.4 was prepared using acidic macromolecules from either an aragonite or calcite layer. These assemblages induced aragonite or calcite formation respectively [Falini et al., 1996].

A large number of proteins are involved in nacre biomineralisation and several have been identified and studied. Recent examples include “calcite blocker” proteins AP7 and AP24 (aragonitic proteins numbered by molecular weight in kDa) [Michenfelder et al., 2003; Wustman et al., 2004; Kim et al., 2004b]; the acidic, aspartate-rich matrix protein Pif, directly involved in creating aragonite platelets in nacre [Suzuki et al., 2009; Kröger, 2009]; and the mineralisation-amplifying protein PFMG1 [Perovic et al., 2013]. There has been an explosion in the number of mollusc shell proteins identified since 2008 [Marin et al., 2013], with expected functions including creating a gel or colloidal region for crystallisation, compartmentalisation of the environment to template future microstructure, selective promotion of crystal nucleation and growth, and crystal growth inhibition [Marin et al., 2013].

1.6 n16N

n16 is a family of 108AA (amino acid chain length) “aragonite promoter” proteins [Samata et al., 1999; Collino and Evans, 2008] in nacre. 23 polymorphic variants have been identified, all actively expressed in pearl oyster (*pinctada fucata*) [Nogawa et al., 2012], while homologues of n16 have been found in other molluscs [Gardner et al., 2011; Marie et al., 2012; Montagnani et al., 2011].

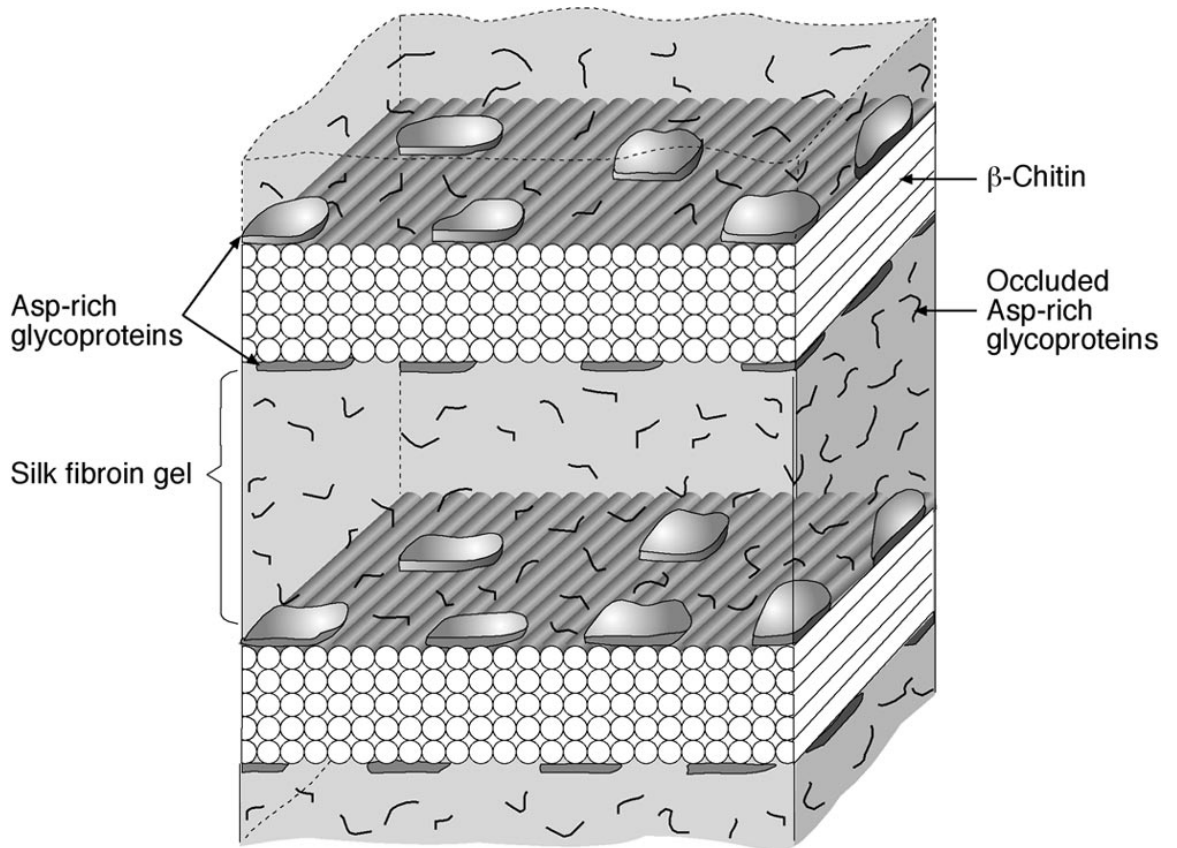


Figure 1.4: [Levi-Kalisman et al., 2001]

A putative scheme for the organic biomaterialization matrix which grows nacre. Interlamellar matrix sheets are composed mainly of aligned β -chitin fibres in several layers, with acidic glycoproteins at their surface which lead to electron-dense patches (not labelled). Aragonite biomaterialisation occurs in the disordered silk fibroin gel region, most likely nucleating epitaxially on the chitin framework [Weiner et al., 1984].

The 30AA N and C terminal regions have been used as n16 mimics for study and have been found to control the morphology of calcium carbonate crystals [Kim et al., 2004a]. The N-terminal sequence of n16 shown in fig. 1.5, named n16N, has been studied in some detail and has been called “the key self-assembly/aragonite forming domain” [Seto et al., 2014].

n16N was found to select the aragonite polymorph when adsorbed on β -chitin [Keene et al., 2010a], while the addition of silk fibroin from silkworm (*bombyx mori*) resulted in metastable vaterite and amorphous calcium carbonate (ACC) [Keene et al., 2010b]. Sequence-scrambling and point mutations to replace the acidic residues with neutral residues (asp \rightarrow asn, glu \rightarrow gln) greatly hampered the selectivity of the peptide in both studies. It is confirmed that Asp and Glu do have an active role in organic-mineral association [Metzler et al., 2008] and that these substitutions abolish n16N’s ability to form complexes with Ca^{2+} [Delak et al., 2007].



Figure 1.5: Amino acid sequence of the 30AA N-terminal region of n16, called n16N. An ellipsis indicates where the full n16 sequence continues, and braces indicate suggested subdomains [Brown et al., 2014], summarised in table 1.3. Cationic amino acid residues shown in bold blue, anionic residues shown in bold red. The last 14 residues, labelled SD3, represent a highly charged region which may be the mineral assembly subdomain.

Further studies shed some light on the mechanism involved. n16N assembles to form fibril-spheroidal oligomers which retain a high level of secondary structural disorder similar to the monomeric state. Amorphous sheet or film assemblies can also form. These complexes act as nucleation sites for crystal growth; fig. 1.6 shows an *in vitro* case of pre-formed assemblies of n16N exhibiting polymorph selection. The assembly process is pH dependent and does not require Ca^{2+} ions, although their presence does shift assembly to a higher required pH [Amos et al., 2011].

n16N neither binds strongly to added Ca^{2+} ions [Seto et al., 2014], nor significantly changes its conformational ensemble [Collino and Evans, 2008; Brown et al., 2014] in their presence. However, there is recent suggestion that n16N assemblies create localised compartments of Ca^{2+} , in which vateritic pre-nucleation clusters are stabilised [Seto et al., 2014]. Vaterite is a precursor to aragonite [Bischoff, 1968].

n16N subdomains with the capacity to perform different functions, especially in the context of the three-component n16N/ β -chitin/calcium carbonate system, have been proposed [Brown et al., 2014]. These are described in table 1.3 and

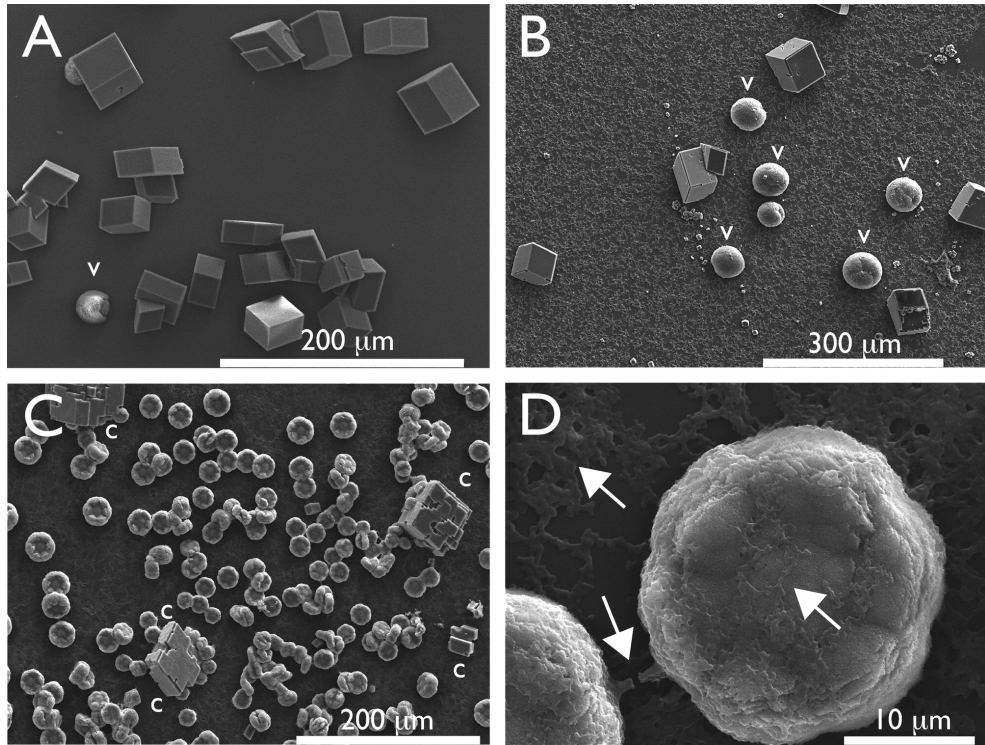


Figure 1.6: [Amos et al., 2011]

Two-stage mineralisation experiment to test whether pre-formed n16N assemblies could nucleate calcium carbonate polymorphs. In stage one, n16N deposits are allowed to form in typical mineralisation conditions (50 μ M n16N, 16h, 16 $^{\circ}$ C). In stage two, washed supports with n16N deposits are transferred to a mineralisation solution with or without n16N content.

(A) *Calcium carbonate mineralisation without n16N IDP. Calcite polymorph dominates.*

(B) *n16N present during first stage of mineralisation, yielding more vaterite (v).*

(C,D) *n16N present during both stages of mineralisation. The arrows indicate fibril-spheroidal deposits of n16N on the exposed surface of a crystal.*

shown in fig. 1.5. SD1 and SD2 are rich in tyrosine, which are hypothesised to have roles in intra- and inter-peptide stabilisation via ring-ring and hydrogen bond interactions, and lead to SD1 and SD2 being less flexible than SD3 [Brown et al., 2014].

Name	Residue indices	Notes
SD1	1 to 8	Tentative role in Y-mediated β -chitin binding. Intrapeptide stabilisation.
SD2	9 to 16	Clustering role due to interpeptide Y-Y interactions [Evans, 2012]. Intrapeptide stabilisation.
SD3	17 to 30	Greatest conformational flexibility, highly charged; proposed “fly-casting” mechanism in ion capture [Shoemaker et al., 2000]

Table 1.3: *Suggested roles of the subdomains of n16N [Brown et al., 2014].*

1.7 Summary

This chapter lays out that intrinsically disordered proteins are a recently emerging research field of great interest due to the novel functions and mechanisms that these proteins fulfill and use. n16N is one of myriad biomineralisation proteins, exhibiting polymorph selectivity on calcium carbonate in solution. Several experimental papers have been published on the activity of n16N in a range of assay conditions, but the effort to understand n16N and other IDPs could be augmented by accelerated-sampling simulation.

Chapter 2

Accelerated simulation techniques

A fundamental goal of the project as laid out in the beginning of chapter 1 is to employ *accelerated simulation techniques*. These are techniques which reduce the computational load involved in characterising a simulated system. Such techniques are grouped into *coarse-graining*, which simplify the representation of molecules of a system, and *accelerated sampling*, which alter the dynamics to characterise the system's phase space in fewer (pseudo-)time-steps.

This chapter will give some background to acceleration techniques, and name the choices of accelerated techniques to be used in the project.

2.1 Coarse-graining

Coarse-graining in computer simulation refers to any scheme of representation in which the basic unit of simulation is greater than a single atom. Integrating out a system's degrees of freedom through coarse-graining can make feasible simulation times rise by orders of magnitude, lifting the existing limit of tens or hundreds of nanoseconds for an atomistic protein in solvent. A coarse-grained simulation has a smoother energy landscape, and thus travels through configurational space faster [Molinero and Goddard, 2004]. This improves the sampling of conformations, but blurs the already questionable connection between simulation time and real time.

The choice of coarse-graining scheme is arrived at by deciding both the mapping of real atoms to groups of united pseudo-atoms, and defining the potential U_{CG} , with which the protein interacts. These decisions are influenced by the goal of the scheme: schemes may set out to be transferable, or relevant only to certain

problems.

2.1.1 Defining the CG mapping

A mapping may have between less than one bead per amino acid to six or so. Single-bead models tend to fall into a class of ‘Gō-like’ models, in which a bias towards a reference native configuration exists; this is necessary to compensate for the very crude description of each amino acid. The usefulness of this very simple approach is in the fact that it recreates the funnel towards the native state in the energy landscape, which, for ordered proteins, must be present in the actual protein’s landscape too [Baker, 2000].

Models of four or so beads generally either focus on detail in the side-chain, or detail in the back-bone. The MARTINI forcefield and others fall into the former category [Hills et al., 2010; DeVane et al., 2009; Monticelli et al., 2008]. The simplicity of the back-bone is dealt with by fixing the secondary structure, which makes the model inadequate to model conformational changes of secondary structure. By the same token, this is not a suitable coarse-graining scheme for IDPs. On the flipside, models which give an almost all-atom description of the back-bone by modelling each of C_α , C' and N explicitly can aim at organically finding secondary structure, thus offering a better hope of understanding IDPs [Bereau and Deserno, 2009; Barducci et al., 2011; Coluzza, 2011].

Defining the CG potential

The first step of forming the CG potential is to assume that it can be described by pairwise interactions in a predetermined functional form. A CG potential may be either continuous or discontinuous, the latter being far simpler, but able to take advantage of faster discontinuous molecular dynamics [Alder and Wainwright, 1959; Rapaport, 1978; Bellemans et al., 1980]. The PRIME model [Voegler Smith and Hall, 2001; Cheon et al., 2010] is one discontinuous potential CG scheme which will be discussed in more detail below (section 2.1.4).

A typical form for a continuous potential may be [Rader, 2010]:

$$U_{\text{CG}} = U_{\text{bonded}} + U_{\text{non-bonded}}, \quad (2.1)$$

$$U_{\text{bonded}} = \sum_{\text{bonds}} \frac{1}{2} k_r (r - r_0)^2 + \sum_{\text{angles}} \frac{1}{2} k_\theta (\theta - \theta_0)^2 + \sum_{\text{dihedrals}} k_\phi [1 + \cos(n\phi + \phi_0)], \quad (2.2)$$

$$U_{\text{non-bonded}} = \sum_{i \neq j} \left[\frac{q_i q_j}{4\pi\epsilon r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right]. \quad (2.3)$$

In this example, the three bonded terms of equation (2.2) give 6 types of free parameters: $k_r, k_\theta, k_\phi, r_0, \theta_0$ and ϕ_0 . The non-bonded terms of equation (2.3) give 4 more types: q_i, q_j, A_{ij}, B_{ij} . Note that q_i and q_j are free parameters which aim to match the system’s Coulombic interactions. Equation (2.2)’s parameters must be defined for each type of bond, angle, and dihedral. In equation (2.3), q_i and q_j will be properties of types of pseudo-atoms. The terms involving A_{ij} and B_{ij} are Lennard-Jones potentials used to match van der Waals interactions, and these parameters must be defined for each pair of pseudo-atom types.

An abundance of other options for these functional forms exist. For example, it has been suggested that using the Morse potential instead of the Lennard-Jones potential would allow greater time-steps, due to a less steep functional form, without a significant loss of realism [Chiu et al., 2010]. Naturally, extra functional forms to account for other features such as hydrogen bonding will bring their own parameters which need to be defined.

There are two classes of techniques for deriving these parameters. Simulation-derived parameters get the fit from more detailed simulations. Within this class, iterative Boltzmann inversion [Reith et al., 2003] takes a simple starting guess for the potential as a seed and iteratively improves it by comparing the radial distribution function $g(r)$ of the atomistic case with that of the coarse-grained case, and adding an improvement term [Milano and Müller-Plathe, 2005]:

$$V_{j+1}(r) = V_j(r) + k_B T \ln \frac{g_j(r)}{g_{\text{Atomistic}}(r)}, \quad (2.4)$$

where V_j is the j^{th} iteration of the coarse-grained potential. Boltzmann inversion is from the family called inversion methods, which ask the question, “What potential

gives rise to the observed properties?”

Force matching (FM) [Zhou et al., 2007] is another simulation-derived technique worth mentioning. The method seeks to minimise the residual χ^2 describing the difference between atomistic forces (projected onto CG sites) and the coarse-grained forces:

$$\chi^2 = \sum_l^L \sum_m^M |F_{lm}^{\text{Atomistic}} - F_{lm}^{\text{CG}}|^2, \quad (2.5)$$

in which F_{lm} is the force acting upon the m th projected or coarse-grained pseudo-atom out of a total of M such sites, in the l th configuration out of L total all-atom trajectory configurations.

Data derived potentials, better known as knowledge-based potentials, use structural statistics from the Protein DataBank [Berman et al., 2000], and either use the native state as a reference to make a Gō-like model, as previously discussed, or assume a Boltzmann distribution of the potential with a maximum of probability in the native conformation [Rader, 2010]. Although knowledge-based potentials are common, they have been criticised on several points [Ben-Naim, 1997], and recently defended [Mullinax and Noid, 2010].

2.1.2 Explicit and implicit solvation

It is neither useful nor common to coarse-grain a protein or polypeptide under study, only to put it in bulk solvent of high detail. Two superior alternatives are either to coarse-grain the water, and continue to represent it explicitly, or to represent it implicitly through its averaged effects.

The recently developed mW water model [Molinero and Moore, 2009; DeMille and Molinero, 2009] reproduces the structure of aqueous ionic solutions without electrostatic interactions, and overcomes the long-standing inability to simultaneously reproduce the structure and energetics of water, through the introduction of tetrahedral interactions. This model has been used in DNA simulation with some success [DeMille et al., 2011] and the authors expect more uses in biomolecules [Molinero and Moore, 2009]. Another recent model termed WAT FOUR [Darré et al., 2010] offers long-range electrostatics and its own dielectric permittivity (rather than a preset constant value), and advertises a good representation of the aqueous environment in the most biologically relevant temperature range, from 278K to 328K.

Implicit solvent models treat water as a continuum solvent. We assume that the total potential energy of a biomolecule in solvent is decomposable into three

terms [Roux and Simonson, 1999], so that:

$$U(\mathbf{X}, \mathbf{Y}) = U_u(\mathbf{X}) + U_v(\mathbf{Y}) + U_{uv}(\mathbf{X}, \mathbf{Y}) . \quad (2.6)$$

Here, u represents the biomolecule with coordinates \mathbf{X} and v the solvent with coordinates \mathbf{Y} . U_u is the biomolecule's potential, U_v is the potential of the solvent, and U_{uv} is the biomolecule-solvent potential. The ensemble average of any property $\langle Q \rangle$ of the biomolecule depends on the probability $P(\mathbf{X}, \mathbf{Y})$ of the state:

$$\langle Q \rangle = \int d\mathbf{X} d\mathbf{Y} Q(\mathbf{X}) P(\mathbf{X}, \mathbf{Y}) , \quad (2.7)$$

$$P(\mathbf{X}, \mathbf{Y}) = \frac{e^{-U(\mathbf{X}, \mathbf{Y})/k_B T}}{\int d\mathbf{X} d\mathbf{Y} e^{-U(\mathbf{X}, \mathbf{Y})/k_B T}} . \quad (2.8)$$

With equation (2.6), equations (2.7) and (2.8) can be rewritten *without* explicit reference to the solvent degrees of freedom \mathbf{Y} , which are *integrated out*:

$$\langle Q \rangle = \int d\mathbf{X} Q(\mathbf{X}) P(\mathbf{X}) , \quad (2.9)$$

$$P(\mathbf{X}) = \int d\mathbf{Y} P(\mathbf{X}, \mathbf{Y}) , \quad (2.10)$$

$$P(\mathbf{X}) = \frac{\int d\mathbf{Y} e^{-[U_u(\mathbf{X})+U_v(\mathbf{Y})+U_{uv}(\mathbf{X}, \mathbf{Y})]/k_B T}}{\int d\mathbf{X} d\mathbf{Y} e^{-[U_u(\mathbf{X})+U_v(\mathbf{Y})+U_{uv}(\mathbf{X}, \mathbf{Y})]/k_B T}} . \quad (2.11)$$

Equation (2.11) can be simplified with recourse to the potential of mean force $W(\mathbf{X})$ [Kirkwood, 1935]:

$$P(\mathbf{X}) = \frac{e^{-W(\mathbf{X})/k_B T}}{\int d\mathbf{X} e^{-W(\mathbf{X})/k_B T}} . \quad (2.12)$$

Therefore, an effective potential $W(\mathbf{X})$ exists which preserves all information about the influence of the solvent on the equilibrium properties of the biomolecule [Roux and Simonson, 1999]. This is the goal of an implicit solvation model.

In proteins, residues may be attracted to each other due to hydrophobicity, and this can be incorporated into non-bonded interaction terms between residues as a means of capturing the solvent's effects implicitly [Bereau and Deserno, 2009].

2.1.3 Validating a CG model

There is no obvious objective means with which to compare CG model quality. However, techniques exist which can provide some ranking.

Decoy sets [Samudrala and Levitt, 2000] are large sets of conformations of a given protein, in which one conformation is the true native state, and the others are decoys. A CG model can be tested on its ability to distinguish the native structure from the decoys, and can be given a score based on how many or few decoys ‘fooled’ the model. Many sets relevant to one application may be used, to provide a reliable and pertinent result [DeVane et al., 2009].

A more recent approach is to use the relative entropy S_{rel} of a model system, given by

$$S_{\text{rel}} = \sum_i p_{\text{T}}(i) \ln \frac{p_{\text{T}}(i)}{p_{\text{M}}(i)}, \quad (2.13)$$

in which $p_{\text{T}}(i)$ is the probability of configuration i in the target system, and $p_{\text{M}}(i)$ is that of the model system. This is argued by Shell [Shell, 2008; Chaimovich and Shell, 2010] to be of fundamental physical significance to multiscale problems, with a minimum in this function representing a minimum in information lost in coarse-graining.

2.1.4 Choices of coarse-grained models

Two coarse-grained models have been selected as suitable for the goals of the project. These are the continuous-potential PLUM model [Bereau and Deserno, 2009] and the discontinuous-potential PRIME20 model [Cheon et al., 2010]. fig. 2.1 shows that both models feature a four-site-per-residue level of coarse-graining, with three sites representing the backbone and a single bead for the side-chain.

The models are designed for implicit solvation and aim to natively find secondary structure. Neither model explicitly includes charged interactions, though this information is partially included in the resultant parameters, because they come from statistical analyses of data on atomistic and residue distance distributions. Both include directional interaction potentials for backbone-to-backbone hydrogen bonding. Both models’ potentials are knowledge-based for the crucial side-chain non-bonded interactions.

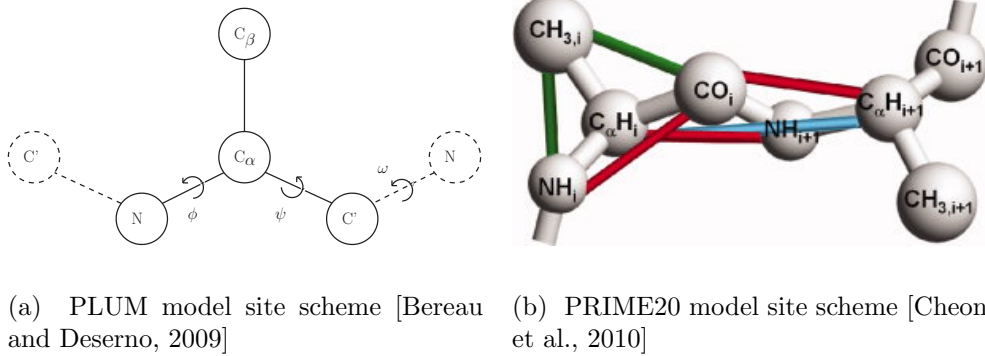


Figure 2.1: *The PRIME20 and PLUM models are identical in their mappings of atoms to coarse-grained sites. All residues feature a single side-chain site, except for glycine which has no side-chain site. The PLUM model backbone maintains its spatial arrangement through continuous-potential bonds, angles and dihedrals as in equation (2.2), while the PRIME20 model relies upon square bonds and square pseudo-bonds, seen in fig. 2.1b.*

PLUM

The PLUM model, introduced in 2009 [Bereau and Deserno, 2009], is designed to model protein folding and aggregation, with a goal of being useful where the protein’s conformation is “not known, not well defined, strongly perturbed from the native state, or adjusts during aggregation events”.

Non-bonded side-chain chain interactions take the form:

$$U_{\text{hp}}(r) = \begin{cases} 4\epsilon_{\text{hp}} \left[\left(\frac{\sigma_{C_\beta}}{r} \right)^{12} - \left(\frac{\sigma_{C_\beta}}{r} \right)^6 \right] + (\epsilon_{\text{hp}} - \epsilon'_{ij}), & r \leq r_c, \\ 4\epsilon_{\text{hp}} \epsilon'_{ij} \left[\left(\frac{\sigma_{C_\beta}}{r} \right)^{12} - \left(\frac{\sigma_{C_\beta}}{r} \right)^6 \right], & r_c \leq r \leq r_{\text{hp,cut}}, \\ 0, & r > r_{\text{hp,cut}}, \end{cases} \quad (2.14)$$

with the Van der Waals side-chain bead radius $\sigma_{C_\beta} = 5.0 \text{ \AA}$, normalised interaction strengths ϵ'_{ij} , mixed for residue type i and j as $\epsilon'_{ij} = \sqrt{\epsilon'_i \epsilon'_j}$ and free parameter ϵ_{hp} .

The ϵ_i values are derived from a study of protein crystal structures [Miyazawa and Jernigan, 1996], which resulted in a 20×20 table of relative interaction strengths ϵ_{ij} . These were reduced in the PLUM work to 20 interaction parameters ϵ_i which are designated as hydrophobic interactions U_{hp} , but also capture some other effects. The process of reduction was referred to by the authors as “deconvolution”. The mathematical processes involved were not made clear, but the original parameters ϵ_{ij} and the new quantities ϵ_i have a correlation coefficient of 98%.

PRIME20

The PRIME model was originally introduced in 2001 [Voegler Smith and Hall, 2001] to study secondary structure formation. [Cheon et al., 2010] document efforts to extend the PRIME coarse-grained protein model to be usable for a protein made of any of the 20 amino acids. Unlike Bereau’s scheme, the PRIME20 model uses discontinuous potentials in the form of square wells and hard spheres. This is an austere coarse-graining of the potential, but, if this approach is successful in finding secondary structure, it has a big advantage in its ability to employ discontinuous molecular dynamics [Alder and Wainwright, 1959; Rapaport, 1978; Bellemans et al., 1980].

PRIME20 uses a knowledge-based potential (see section 2.1.1); a set of 711 protein structures which were not membrane proteins or multi-domain proteins, did not have missing atoms, partial representations, broken chains, or non-standard amino acids, and did not feature ligands, small single helices or coiled peptides was obtained from the Protein DataBank [Berman et al., 2000] for use in parametrisation. The three pseudo-atoms NH, $C_\alpha H$, and CO were mapped to the centres of the backbone atoms N, C_α , and C, respectively, and the single side-chain pseudo-atom was mapped to the centre of mass of the PDB side chain. In this way, distribution functions for all pairs of atoms were plotted. Instead of regular radial distribution functions, which provide no obvious choice for well radius, a distribution function is used which only counts pairs of united atoms in which more than half of the distances between the heavy atoms of the first united atom and that of the second united atom are less than 5.5 Å. The justification for picking 5.5 Å is not made clear. The sphere diameters are then chosen to be equal to the lowest distance for which the distribution function is nonzero, and the well radii are chosen to capture approximately 90% of the distribution.

The crucial non-bonded side-chain interactions are more sophisticated in PRIME20 than in PLUM in two ways. First, bead sizes and bead well diameters for the side-chain bead C_β are arrived at separately for each residue in PRIME20, whereas in PLUM a single Van der Waals radius of 5.0 Å is used. Secondly, PRIME20’s non-bonded interactions are parameterised for *pairs* of side-chain site types, rather than for *single* side-chain sites which then require mixing rules. This allows for differences between different amino acids which wouldn’t survive a mixing scheme, such as charge and side-chain hydrogen bonding, to be implicitly represented. The non-bonded side-chain well-depth parameters ϵ_i are estimated using a stochastic iterative learning algorithm, which seeks to ensure that generated decoy structures are less stable than native structures.

The model was utilised by the group originally to fold alanine chains [Voelger Smith and Hall, 2001] and subsequently to form secondary and tertiary β -structure from short chains in multiplicity with positive results [Cheon et al., 2011; Wagoner et al., 2012], however, third-party testing is lacking.

2.2 Accelerated sampling techniques

Accelerated sampling techniques seek to sample a system’s configurational space efficiently. The properties of the system are altered in unphysical ways to this end, with the caveat that the system’s properties of interest must be recoverable. Accelerated sampling techniques fall into 3 broad categories [Laio and Gervasio, 2008].

The transition mechanism group of acceleration techniques seek to sample in a manner useful to understanding the kinetics of a transition. For example, in transition path sampling, paths are generated by importance sampling of trajectory space, such that most time is spent sampling transition paths [Bolhuis et al., 2002]. Developing a coarse-grained model which displays realistic transition properties would be a difficult task. Furthermore, an *a priori* knowledge of the two states between which a transition of interest will be studied is required [Juraszek and Bolhuis, 2006]. Since understanding the transition paths in detail is not a key part of this project, there is no reason to invest time in these methods.

Techniques in which *collective variables*, thought to characterise the macrostate of the system, are picked, and aim to flatten the probability distribution with respect to these collective variables, make up the second family. For example, in metadynamics [Laio and Parrinello, 2002], the system is forced out of local minima in the free energy landscape by adding terms to the potential which are Gaussian in collective variable-space [Laio and Gervasio, 2008]. The choice of collective variables must be made with care; a poor choice will not aid the full exploration of accessible phase space and can mask the true nature of the energy landscape. A popular choice for a collective variable in the case of phase transitions is the potential energy [Trudu et al., 2006]. It is not clear or obvious what a good choice would be for IDPs.

The third family of techniques makes use of high temperatures to explore phase space efficiently alongside the lower temperatures which are actually of interest. Alternatively, a similar effect is achieved by exploring phase space in a biased manner, as a function of the potential energy. These methods are quite generally applicable, and therefore will be used in this project. Two techniques from this family will be described in detail and used.

2.2.1 Replica exchange molecular dynamics

Replica exchange molecular dynamics [Sugita and Okamoto, 1999] is a reformulation of replica exchange Monte Carlo, which was developed by Swendsen and Wang [1986] and allows low-temperature simulations to easily explore phase space, with reduced hindrance from potential energy barriers.

M non-interacting systems indexed $i \in \{1, \dots, M\}$, with the same Hamiltonians \mathcal{H} , but in different heat baths at temperatures T_i obeying $T_i < T_{i+1}$, attempt to swap states with their nearest-temperature neighbours. In the Monte Carlo formulation, only the coordinates \mathbf{X}_i had to be swapped. In molecular dynamics, rather than swapping the momenta \mathbf{P}_i , the current momenta are scaled according to the new heat bath's temperature $T_{i\pm 1}$ and the old heat bath's temperature T_i :

$$\mathbf{P}_i \rightarrow \sqrt{\frac{T_{i\pm 1}}{T_i}} \mathbf{P}_i. \quad (2.15)$$

This is a natural choice in order to satisfy the average kinetic energy at temperature T , in a system with N particles [Sugita and Okamoto, 1999]:

$$\langle E_K \rangle_T = \left\langle \sum_k^N \frac{\mathbf{P}_k^2}{2m_k} \right\rangle_T = \frac{3}{2} N k_B T. \quad (2.16)$$

The detailed balance condition can be satisfied by the usual Metropolis prescription Sugita and Okamoto [1999], leading to a transition matrix specified by:

$$T(j \leftarrow k) = \begin{cases} \alpha(j \leftarrow k), & \rho_j \geq \rho_k, j \neq k; \\ \alpha(j \leftarrow k) \frac{\rho_j}{\rho_k}, & \rho_j < \rho_k, j \neq k; \\ 1 - \sum_{j \neq k} T(j \leftarrow k), & j = k; \end{cases} \quad (2.17)$$

$$\rho = \exp \left\{ - \sum_i^M \beta_i \mathcal{H}(\mathbf{X}_i, \mathbf{P}_i) \right\}. \quad (2.18)$$

Here, k is the current state of the generalised ensemble, and j is the state after an exchange. Move proposal probability is $\alpha(j \leftarrow k) = \alpha(k \leftarrow j)$ and Boltzmann weight for the generalised ensemble is ρ . Following this prescription, swap moves do not disturb the Boltzmann distribution corresponding to any particular canonical ensemble [Frenkel and Smit, 2002, p. 389]. Resultantly, ensemble averages can be determined as in a regular simulation, but with far less computation required to

adequately sample phase space.

A major drawback of this method is that the canonical sampling means that unlikely states may not be visited at all, so the estimated free energy may have large errors in some regions.

Due to the acceptance criteria and the nature of the Boltzmann weighting in equation (2.18), the acceptance probability declines exponentially with the difference between the two inverse temperatures β Sugita and Okamoto [1999]. Furthermore, the number of replicas required for efficient sampling has been shown to scale as $\mathcal{O}(f^{1/2})$, f being the number of degrees of freedom of the system [Fukunishi et al., 2002]. Clearly, this is more problematic in an explicitly solvated simulation, in which f is greater. To counteract this, an altered method, known as Replica Exchange with Solute Tempering or REST [Liu et al., 2005] and since developed into REST2 [Wang et al., 2011] has been created. REST2 achieves a scaling of $\mathcal{O}(x)$, where $x \geq f_s$, for a number of degrees of freedom of the solute alone f_s [Wang et al., 2011]. This is achieved with a deformed Hamiltonian, causing the solute-solute and solute-solvent interactions to be tempered, while solvent-solvent interactions are unaffected and occur at the reference temperature.

2.2.2 Statistical temperature molecular dynamics

Statistical temperature molecular dynamics (STMD) [Kim et al., 2006, 2007] is a flat-histogram sampling technique which is based upon the Wang-Landau Monte Carlo algorithm [Wang and Landau, 2001b,a]. In Wang-Landau sampling, the central idea is to arrive at a flat potential energy distribution by weighting the Monte Carlo acceptance probability by $w(U) = 1/\Omega(U)$. $\Omega(U)$ is the density of states, connected to the microcanonical entropy as $S(U) = k_B \ln \Omega(U)$. This leads to a uniform random walk in potential energy space.

$\Omega(U)$ is not known *a priori*, so a running estimate $\tilde{\Omega}(U)$ is kept. Initially, $\tilde{\Omega}(U) = 1$ is set. Next, every time an energy U_i is visited, the density of states is updated via the operation $\tilde{\Omega}(U_i) \rightarrow f\tilde{\Omega}(U_i)$, where f is a modification factor and is greater than 1. Each update operation diminishes the probability of a return visit to U_i .

A histogram of visits to energy states is kept, and, when it becomes flat within a given tolerance, the modification factor f is decreased (conventionally by the operation $f \rightarrow \sqrt{f}$), the histogram is reset, and the simulation continues. In the limit $f \rightarrow 1$, $\tilde{\Omega}(U) \rightarrow \Omega(U)$.

In STMD, the statistical temperature $\tilde{T}(U)$ is the object of the running estimate, instead of $\tilde{\Omega}(U)$. The two are connected as follows:

$$\frac{1}{\tilde{T}(U)} = \frac{\partial \tilde{S}(U)}{\partial U} = k_B \frac{\partial \ln \tilde{\Omega}(U)}{\partial U}. \quad (2.19)$$

Rewriting as a central finite difference approximation:

$$\frac{1}{\tilde{T}(U_i)} \approx \frac{\tilde{S}(U_{i+1}) - \tilde{S}(U_{i-1})}{2\Delta}, \quad (2.20)$$

in which Δ is the difference in energy between two adjacent bins in the statistical temperature histogram. Feeding the WL update scheme $\tilde{\Omega}(U_i) \rightarrow f\tilde{\Omega}(U_i)$ into the microcanonical entropy $S = k_B \ln \Omega(U)$, we obtain an entropy update scheme $\tilde{S}(U_i) \rightarrow \tilde{S}(U_i) + k_B \ln f$. Equation (2.20) shows that the statistical temperature estimate $\tilde{T}(U_i)$ needs to be updated if either $U_{i\pm 1}$ are visited. Given that the current value for $\tilde{T}(U_i)$ is derived from equation (2.20), the entropy update scheme tells us how to update $\tilde{T}(U_i)$ if U_{i+1} is visited;

$$\frac{1}{\tilde{T}_{\text{new}}(U_i)} \approx \frac{\tilde{S}_{\text{new}}(U_{i+1}) - \tilde{S}(U_{i-1})}{2\Delta} = \frac{1}{\tilde{T}_{\text{old}}(U_i)} + \frac{k_B \ln f}{2\Delta}, \quad (2.21)$$

or if U_{i-1} is visited;

$$\frac{1}{\tilde{T}_{\text{new}}(U_i)} \approx \frac{\tilde{S}(U_{i+1}) - \tilde{S}_{\text{new}}(U_{i-1})}{2\Delta} = \frac{1}{\tilde{T}_{\text{old}}(U_i)} - \frac{k_B \ln f}{2\Delta}. \quad (2.22)$$

Equivalently, if U_i is visited, both $\tilde{T}(U_{i\pm 1})$ must be updated as follows:

$$\frac{1}{\tilde{T}_{\text{new}}(U_{i\pm 1})} \approx \frac{1}{\tilde{T}_{\text{old}}(U_{i\pm 1})} \mp \frac{k_B \ln f}{2\Delta}, \quad (2.23)$$

and this is the STMD update scheme.

An alternative but equivalent formulation due to Allen and Quigley [2013], which highlights the similarity between STMD and Wang-Landau further, is to preserve the Wang-Landau entropy update scheme, $\tilde{S}(U_i) \rightarrow \tilde{S}(U_i) + k_B \ln f$, and to calculate $\tilde{T}(U_i)$ when needed as a central difference approximation, using equation (2.20).

Either of these can be readily implemented into statistical temperature *Monte*

Carlo, altering the acceptance probability to obtain non-Boltzmann sampling. To implement statistical temperature in *molecular dynamics*, the generalized ensemble simulation technique is used [Nakajima et al., 1997], with the temperature held at T_0 . In the canonical ensemble,

$$P(U) = \frac{1}{Z} \Omega(U) e^{-U/k_B T_0}, \quad (2.24)$$

in which Z is the canonical partition function and T_0 is the thermostat-maintained temperature of the simulation, and not the statistical temperature estimate. A flat distribution is obtained by altering the potential U to be $U_{\text{MC}}(U)$:

$$\Omega(U) e^{-U_{\text{MC}}(U)/k_B T_0} = C, \quad (2.25)$$

where C is a constant of choice; let $C = 1$ for ease. Thus,

$$U_{\text{MC}}(U) = k_B T_0 \ln \Omega(U) = T_0 S(U). \quad (2.26)$$

Implementing this potential gives

$$\mathbf{f}_{\text{STMD}} = -\nabla(T_0 S(U)) = -T_0 \frac{\partial S(U)}{\partial U} \nabla U = \frac{T_0}{T(U)} \mathbf{f}_{\text{True}}, \quad (2.27)$$

where \mathbf{f}_{STMD} is the scaled force on each atom due to the multicanonical potential, and \mathbf{f}_{True} is the force on the respective atom due to the normal (canonical) potential.

After collecting the data on $\tilde{T}(U)$ in a simulation, the estimate for the entropy is then given by

$$\tilde{S}(U) = \int_{U_l}^U \frac{1}{\tilde{T}(U)} dU, \quad (2.28)$$

with an arbitrary lower integration limit U_l . Now the canonical ensemble average of an observable can be calculated for the system.

2.3 Summary

The problem of simulation of intrinsically disordered proteins is a recent one, and it would be of great utility to the intrinsically disordered protein community to develop reliable methods by which simulations of intrinsically disordered proteins can be accelerated. Little is known about the degree to which protein representations can be coarse-grained in the case of IDPs, but it is less likely that models with significant simplifications of the backbone could yield good results. Two models have been selected for experimental study with n16N systems; PLUM and PRIME20, and these both maintain three beads per residue on the backbone, and one on the side-chain.

Accelerated sampling schemes disconnect a molecular dynamics experiment from realistic reproductions of the dynamics by introducing unphysical alterations, leaving the statics of the system available for retrieval. Two very general methods have been recruited here; replica exchange and statistical temperature. The former simulates the system at high temperature in parallel to the reference temperature, and executes Monte Carlo style dice throws to swap the systems' configurations in keeping with canonical ensemble probability statistics. The high temperature replica explores phase space relatively freely. The latter implements a multicanonical potential which progressively reduces the probability of visiting the most likely configurations of the system through learning the density of states $\Omega(U)$. The system arrives at a uniform sampling of the potential energy space, and data from this uniform sampling regime can be reconstructed into canonical statistics.

Chapter 3

Software modifications and validation

The principle technical work which was necessary to advance with the project is described in this section. In section 3.1, the parametrisation of the hard-particle model used in the DynamO program is described, and in section 3.2, the PLUM model's implementation into LAMMPS is described.

3.1 PRIME20 initial parametrisation

The chief barrier to implementation of the PRIME20 model was the absence of many crucial parameters from the literature. The antecedent model, PRIME, lays out the backbone geometry, alanine side-chain bead geometry, and hydrogen bonding geometry and energetics [Voegler Smith and Hall, 2001], which are inherited by PRIME20. A parameter governing the tolerance of bond distance fluctuations was updated in a 2004 paper [Nguyen et al., 2004] to induce a more realistic exploration of (ϕ, ψ) phase space. The full PRIME20 model debuted in 2010 and filled in square-well interaction energies, hard sphere radii and square-well radii for side-chain to side-chain interactions, as well as backbone to side-chain interaction energies [Cheon et al., 2010]. Absent from these papers are backbone to side-chain hard sphere and square-well radii, bond and pseudo-bond lengths attaching the side-chains to the backbone, side-chain bead masses and parameters changed since PRIME.

Fragments of this information are scattered through the later literature [Cheon et al., 2011; Wagoner et al., 2011; Cheon et al., 2012], but the picture remains far from complete. After attempts to communicate with the original authors on this point proved unfruitful, the decision was made to advance by delving into parametri-

sation in order to create a PRIME20-based model.

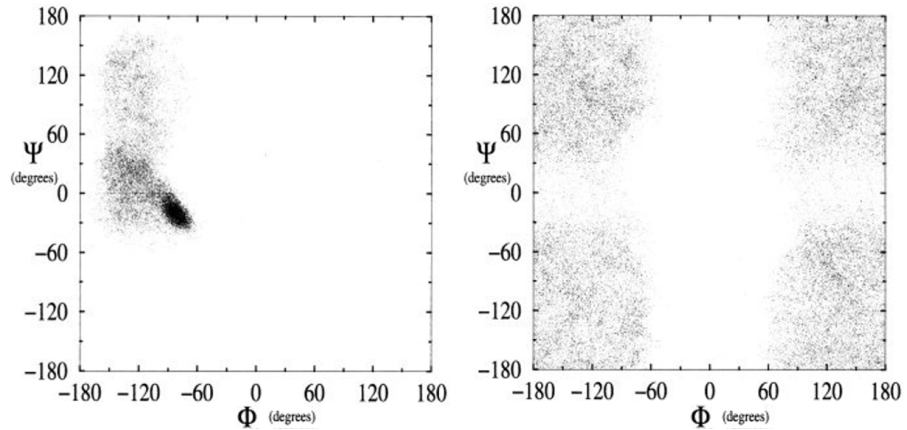
3.1.1 PRIME model validation

The PRIME model can retrospectively be seen as almost a subset of the PRIME20 model; only one side-chain bead type, no non-bonded interactions except for hydrogen bonding, and two modified parameters. Since its description in the literature is complete, it provides an excellent starting point for validation. The original PRIME model will be referred to as PRIME2001, to distinguish it from the larger PRIME20 model. In this section, simulations are run for at least 2.4×10^6 collisions in a large periodic box.

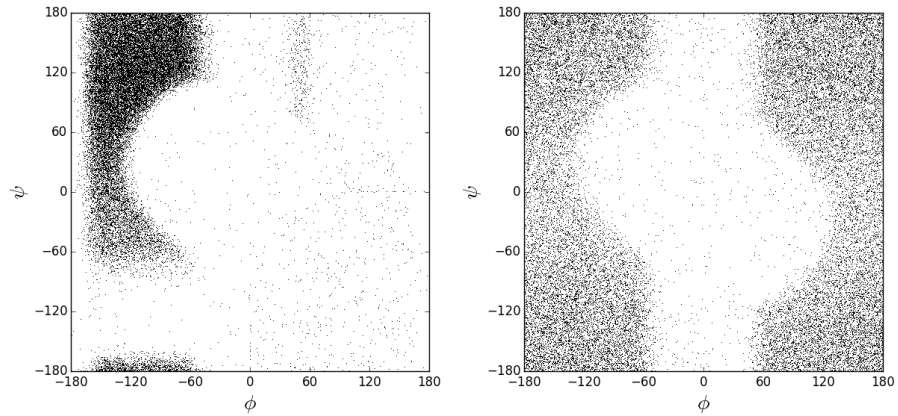
Unfortunately, the parameter set provided as the PRIME2001 model does not lead to the behaviour shown in the reference work [Voegler Smith and Hall, 2001]. The model possesses a very limited ability to hydrogen bond when a side-chain bead is present. With an example system of A20 at a reduced temperature of 0.07 in a REMD simulation, the system averages just 32.70% of the maximum number of possible backbone hydrogen bonds. Looking at α -type backbone hydrogen bonds exclusively (i.e. with a spacing obeying $i + 4 \rightarrow i$), the number is 0.02%. As shown in fig. 3.2, the equilibrated A20 system ought to nearly be a full α -helix all of the time at any temperature below 0.125, according to the reference work.

In 2012 the parameter giving the maximal distance between NH and CO beads for a hydrogen bond to exist was updated from its earlier stated value of 4.20 to 4.50; implementing this change in the current work improved the accessibility of hydrogen-bonded conformations, but did not fully fix the problem. Instead, it was found that the two percentages given above rise to 69.14% and 42.77%, respectively. It is telling that the π -helix, defined by $i + 5 \rightarrow i$, emerged as a common motif, scoring 45.28%. This conformation may be popular because the spiral of the chain is slightly less tightly wound, hinting at an underlying steric clash. (Note that we should not expect consistency between these percentages, because the denominator of the fraction is varying: the maximum number of hydrogen bonds is taken as N when being type-agnostic, as $N - 4$ when considering α -type only, and $N - 5$ for π -type. N is the number of residues in the chain.)

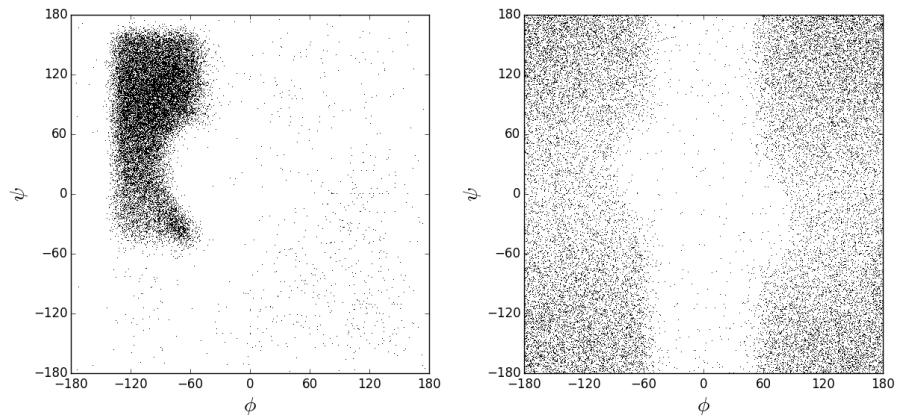
Diagnosis of the problem was made easier through the authors' inclusion of Ramachandran plots, revealing the sterically accessible portion of (ϕ, ψ) phase space. Ramachandran plots in the authors' work do not state the experimental temperature or chain length, but the application of trial and error makes finding a good match possible. In fig. 3.1, the accessible regions in the author's work and the present reproduction are compared.



(a) Reference data; alanine chain [Voegler Smith and Hall, 2001]. (b) Reference data; glycine chain [Voegler Smith and Hall, 2001].



(c) A20 system represented in the canonical PRIME2001 model. (d) G20 system represented in the canonical PRIME2001 model.



(e) A20 system represented in a customised PRIME2001-like model. (f) G20 system represented in a customised PRIME2001-like model.

Figure 3.1: *Ramachandran plots showing the exploration of (ϕ, ψ) phase space for PRIME2001 chains, according to the reference [Voegler Smith and Hall, 2001], our independent reproduction, and a reproduction with a modified parameter set. Note that $T = 0.15$ is a high simulation temperature, certainly above that used by the original authors. By modifying the parameter set, it is possible to reproduce the reference data's accessible regions with moderate accuracy.*

Fig. 1.3 categorised the causes of excluded regions into different steric clashes. This helps identify the wrong-sized clash regions as: **(1)** the central elliptical region due to the $\text{CO}_{i-1}\dots\text{NH}_{i+1}$ clash, which is too large and entirely prohibits the α -helix peak region, **(2)** the lower horizontal region due to the $\text{C}^\beta\dots\text{NH}_{i+1}$ clash, which is too small and erroneously makes values of $\Psi < -45^\circ$ and $\Psi > 170^\circ$ accessible in the A20 simulation, and **(3)** the right vertical region due to the $\text{CO}_{i-1}\dots\text{C}^\beta$ clash, which is too small and erroneously leaves the α_L region prominently accessible.

The beads involved in each clash are implicated in other clashes. Specifically, fixing error **(1)** by reducing the CO bead exacerbates error **(3)**, and doing so by reducing the NH bead exacerbates error **(2)**. These also change the properties of the clashes which are not currently problematic. Many preliminary simulations were carried out to determine empirically that no geometrically consistent set of bead sizes produces matches of figures 3.1a and 3.1b.

Reducing the $\text{CO}_{i-1}\dots\text{NH}_{i+1}$ hard-sphere interaction diameter by 85% caused the central excluded ellipse to shrink to an appropriate size, but jettisoned the geometric consistency of hard-sphere mixing rules. The original authors make use of an allowed 25% overlap between beads separated by less than four bonds along the chain, and our approach to retaining consistency in the model is similar: since the $\text{CO}_{i-1}\dots\text{NH}_{i+1}$ interaction is the only steric consideration with participants separated by as many as four bonds along the chain, we set an allowed 15% overlap between backbone beads separated by this amount. This approach allowed us to respect the author’s original bead sizes for all interactions that do not qualify, and to maintain geometric consistency. This is the sole change between the canonical model of glycine shown in fig. 3.1d and the PRIME2001-like version in fig. 3.1f.

Both errors **(2)** and **(3)** were amenable to an increase in the C^β bead-size. The bead was originally set at a size of 4.408, and an increase to 5.0 was determined by trial and error to be an optimal change. These two alterations combine to produce 3.1e in the simplest possible manner.

It is implied in the source of the model [Voegler Smith and Hall, 2001] that the hydrogen bond well-depth parameter is equal to 1.0 in their system of reduced units. However, a value of 1.26 more accurately reproduces the authors’ plotted behaviour as a function of temperature, and fig. 3.2 shows this. The discrepancy must stem from the differing parameter sets, perhaps the use of a C_β bead at a size of 5.0, which may be large enough to fractionally infringe upon the α -helix’s (ϕ, ψ) -space; see the reference work’s fig. 16. In fig. 3.3, Ramachandran plots are displayed at a temperature causing them to match 3.1a and 3.1b more closely, with regards to peaks and regions which are unfavoured, but not forbidden.

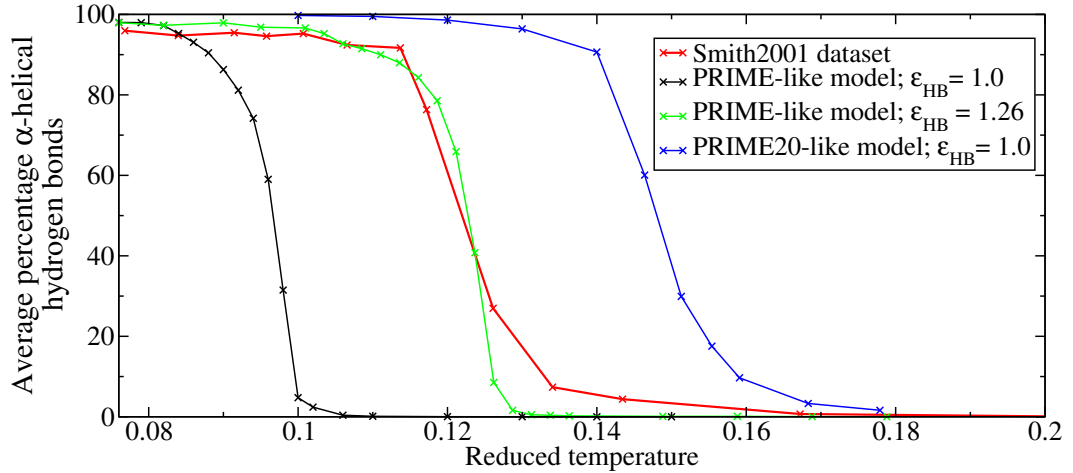
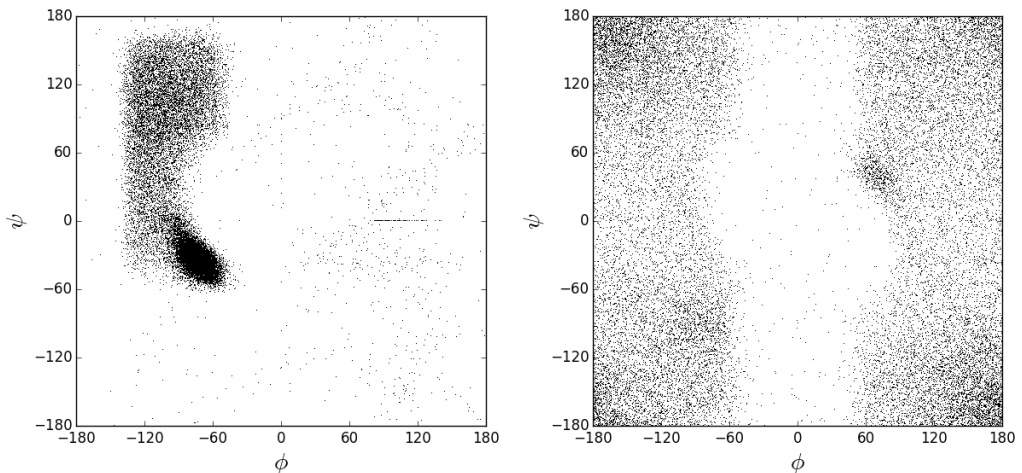


Figure 3.2: Behaviour of a chain of A20 in PRIME2001, PRIME2001-like and PRIME20-like models. After altering the 2001 model to reproduce the authors' accessible areas (see fig. 3.1), we see here that the hydrogen bond interaction strength, ϵ_{HB} , needs raising to 1.26 to match the reference data best. In the PRIME20-like model, α -helices are far more stable. Unfortunately, no data for the behaviour of A20 in the canonical PRIME20 model is available.



(a) A20 system. This is a good qualitative match to fig. 3.1a, though the α -helix peak matching excluded regions to fig. 3.1b, a is centred at $(\phi, \psi) = (-73, -34)$, compared horizontal region centred on $\psi = 0.0$ is less to the source work's $(-82, -19)$.
 (b) G20 system. In addition to having favoured than other allowed regions.

Figure 3.3: Ramachandran plots in the custom PRIME2001-like model at $T = 0.125$, with $\epsilon_{HB} = 1.26$. These plots illustrate that the PRIME2001 data in fig. 3.1a and 3.1b can be reproduced by the custom model.

3.1.2 Parametrising a PRIME20-like model

The PRIME20 model provided non-bonded interaction specifications between side-chain beads, allowing peptides composed of all 20 amino acids to be simulated. Our work with the PRIME20-like model additionally includes an update on the universal bond fluctuation tolerance, which was modified from $\delta = 2\%$ [Voegler Smith and Hall, 2001] to $\delta = 2.375\%$, a value which “produces a more realistic Ramachandran plot” [Nguyen et al., 2004].

Section 3.1 explains that we have been unable to obtain multiple essential parameters from the original authors. The hydrogen bond well diameter has been updated, but its value has been provided subsequently [Cheon et al., 2012]. The side-chain bead masses are missing, but these do not affect ensemble averages. The simple approach used here is to set side-chain bead masses to the sum of the textbook masses of the constituent atoms. Later PRIME20 publications suggest that the same was done for PRIME20 [Cheon et al., 2011; Wagoner et al., 2011; Cheon et al., 2012]. A more nuanced approach which could be useful to study dynamics accurately would be to include mass contributions from bound water. Other solvent effects on dynamics which are not strictly inertial contributions may be best captured via the damping coefficient of a Langevin thermostat, equation (1.11).

This leaves the set of parameters governing backbone to side-chain interactions, including hard sphere diameters, well radii, bond lengths and pseudo-bond lengths. Following the lead of the PLUM model [Bereau and Deserno, 2009], setting these interactions the same as alanine (in the PRIME2001-like model) may be an acceptable simplification, and is certainly more time-efficient than separately parameterising each residue’s set via its impact on the Ramachandran plot.

Therefore, we set bond and pseudo-bond lengths to 2.44 Å, 1.531 Å and 2.49 Å for NH, CH and CO, respectively. We set all non-bonded bead diameters σ_d to values which result from mixing the alanine bead size of 5.0 Å with the PRIME2001-like backbone bead sizes. Well diameters are set somewhat arbitrarily to $1.5\sigma_d$, which brings their ranges to similar values to the non-bonded side-chain to side-chain interaction ranges.

PRIME20 implementation details

The PRIME20-like model was implemented in the event-driven simulation package DynamO [Bannerman et al., 2011]. DynamO is a FOSS (free and open source) program distributed under the terms of the GPL version 3 license [gpl, 2007]. Its source code is written in modern C++11 and is fully object-oriented. The PRIME20-like

model was implemented with great help from the primary developer, Marcus Bannerman. The completed implementation comes to 1356 lines of C++ code. Along with smaller tools to set up simulations in the model, the completed work is now available for the public for use.

The A20 system

The PRIME20 model features an extended range for the hydrogen bond interaction in order to allow for larger side-chains, and the A20 peptide in the PRIME20-like model was found to form far more stable α -helices than the PRIME2001-like model, in which the side-chain is expanded to 5.0 Å, but the hydrogen bond range is not adjusted accordingly. This is plotted in fig. 3.2; unfortunately, no comparable data for the canonical PRIME20 model is known to be available. This result leads to the conclusion that the hydrogen bond interaction strength should be left at 1.0 for the PRIME20-like model.

The 48-peptide $A\beta_{16-22}$ system

An opportunity to compare our PRIME20-like model to the canonical model exists due to a published study using the PRIME20 model on residues 16 to 22 of the β -amyloid ($A\beta$) protein [Cheon et al., 2011], which is linked to Alzheimer’s disease. 48 peptides of the corresponding sequence, KLVFFAE, known as $A\beta_{16-22}$, are placed in a simulation box and aggregation occurs, governed by β -sheet secondary structure.

The original authors simulate the systems at reduced temperatures $T^* \in \{0.1, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, 0.197, 0.20, 0.205\}$, at concentrations 10mM and 20mM, for 5 or 10 runs each.

It was infeasible to recruit a matching level of computational brawn to study this large system in the present project, but trial runs on the system were conducted at $T^* = 0.17$ and $T^* = 0.20$ at 20mM, to check that aggregation occurs as expected and results are similar. Each system ran for 5.0×10^{10} events. Both systems proceeded to aggregate into layered β -sheets, as in the reference work. Fig. 3.4 shows the structures which were observed at the end of each run.

The reference work defines a test of parallel or anti-parallel structure. The orientation of two bound peptides is measured by the angle θ between vectors down each chain, measured from the second C_α atom to the second-from-last. The condition for parallel is $\theta < 60^\circ$, and the condition for anti-parallel is $\theta > 120^\circ$. The current work takes two peptides to be bound in β -structure when they share four or more hydrogen bonds. The simulations carried out here ended with averages of

72.6% and 70.0% anti-parallel structure for $T^* = 0.20$ and $T^* = 0.17$, respectively. These are slightly weaker preferences for anti-parallel structure than the reference work, whose ranges are approximately 90% to 100% for $T^* = 0.20$, and 72% to 90% for $T^* = 0.17$, over five runs.

The original study observes that, rather than perfectly anti-parallel β -sheets coming together when the peptides first aggregate, “parallel strands that formed early in the simulation switch to an anti-parallel orientation by a continuous stochastic process”, and that this is most likely immediately below the fibrilization transition temperature. Therefore, one possible source for the discrepancy in observed β -structure, especially at $T^* = 0.20$, would be a small change in the transition temperature caused by differences from the reference model. Another cause may be the lower run time in the present study; the original study had runs of up to 2.68×10^{11} events, compared to 5.0×10^{10} here.

3.2 Implementation work in LAMMPS

LAMMPS is a FOSS program distributed under the terms of the GPL version 2 [gpl, 1991]. Its source code is written in C++, and the high-level organisational structure is object-oriented, leading to an easily extensible modular code-base. However, all significant computations within each class are performed with low-level C-style data structures and operations, which may be faster.

LAMMPS can be compiled as an executable binary which reads an input script for instructions. Less commonly, LAMMPS can be compiled as a library object, which can then be called and instructed from another program or script. This suggests two paths for modifying and extending LAMMPS.

3.2.1 The PLUM model

Usage of the PLUM model required implementation of the piecewise hydrophobic potential and the 6-body backbone hydrogen bond. The former uses a Lennard-Jones piece in the attractive domain, a Weeks-Chandler-Andersen piece in the repulsive domain, and zero force and potential above the cut-off distance. This enables a consistent excluded volume for each side-chain bead, while the attractive portion is scaled. The hydrogen bond is a 12-10 Lennard-Jones interaction multiplied by a directional term which reaches its minimum when the first residue’s nitrogen-to-hydrogen vector and the second residue’s carbon-to-oxygen vector point at each other [Bereau and Deserno, 2009]. Since backbone hydrogen and oxygen atoms

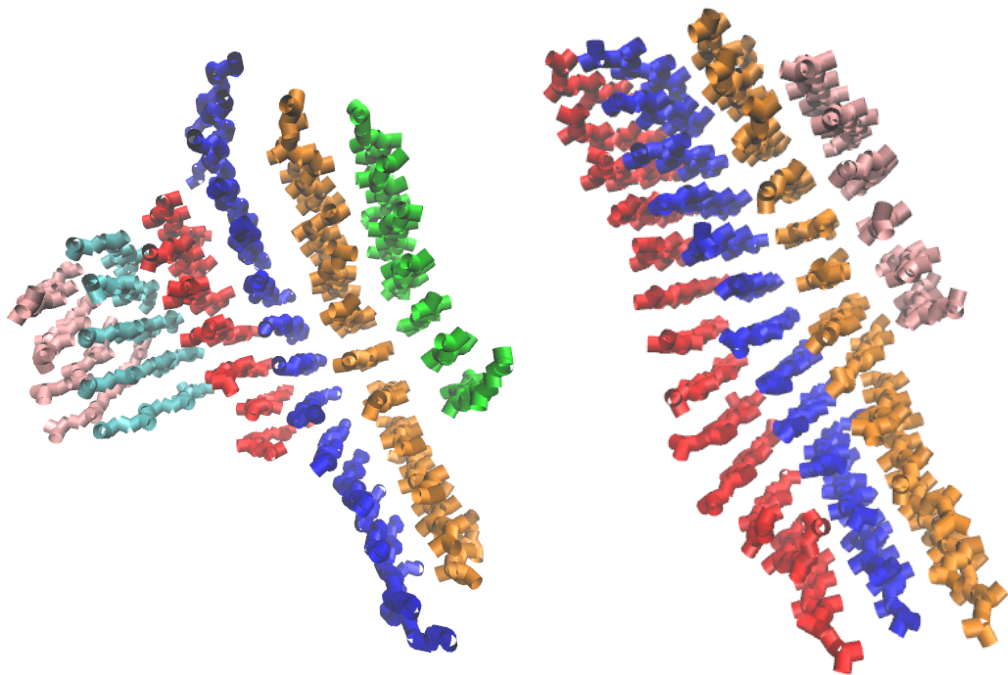


Figure 3.4: *Final structures of the 48-peptide $A\beta_{16-22}$ system in the PRIME20-like model. Both structures form β -sheets, with a preference for anti-parallel β -strands. **Left:** $T^* = 0.17$ simulation. Five sheets feature all 48 chains. **Right:** $T^* = 0.20$ simulation. The 45 chains shown make up four sheets, while one chain is free and two are disordered on the surface of the structure.*

are not explicitly represented, a further complication is calculating their implicit position. The completed implementation comes to 1323 lines of C-style C++ code.

The tripeptide GAG system

Validation begins with replication of the authors’ figure 4b, which shows the free energy landscape of (ϕ, ψ) phase space for the tripeptide GAG. This depends upon the bonded interactions and leaves out the non-bonded interactions and hydrogen bonding. A good match between the original authors’ work and the present work is shown in fig. 3.5.

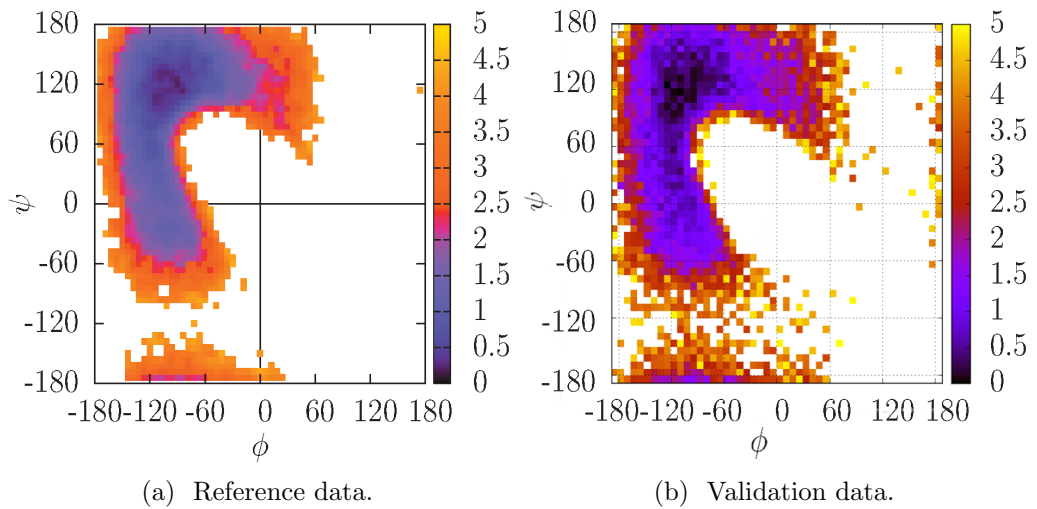


Figure 3.5: *Validation data for the bonded components of the PLUM model, contrasted with the reference data [Bereau and Deserno, 2009]. The free energy landscape of the Ramachandran plot is shown for the GAG tripeptide at $T^* = 1.0$. The colour represents the free energy difference with the lowest conformation in reduced units. Both datasets result from REMD simulations at reduced temperatures 0.5, 0.7, 1.0, 1.3, 1.6, 1.9, 2.2 and 2.5. WHAM [Ferrenberg and Swendsen, 1988; Kumar et al., 1992, 1995] and MBAR [Shirts and Chodera, 2008] are used to produce reweighted analyses in the reference data and the validation data, respectively.*

15-unit GNNQQNY system

For validation of the complete model, we reproduce the authors’ simulation of 15 units of the peptide sequence GNNQQNY. This is an aggregation system which forms parallel β -sheets at low temperatures, transitioning to random coil monomers at higher temperatures. The authors’ use of 8 replicas for a period of 50 ns each yields inconsistent results in our initial simulations, suggesting an under-sampled

simulation. Increasing the sampling to 30 replicas with more in the transition zone and increasing the simulation time to 330 ns leads to consistent results and a far sharper heat capacity peak, remaining in agreement on transition temperature with the original dataset. This is shown in fig. 3.6. As with the reference simulations, at temperatures below the peak, parallel β -sheets are observed, and above the peak, random coil behaviour dominates.

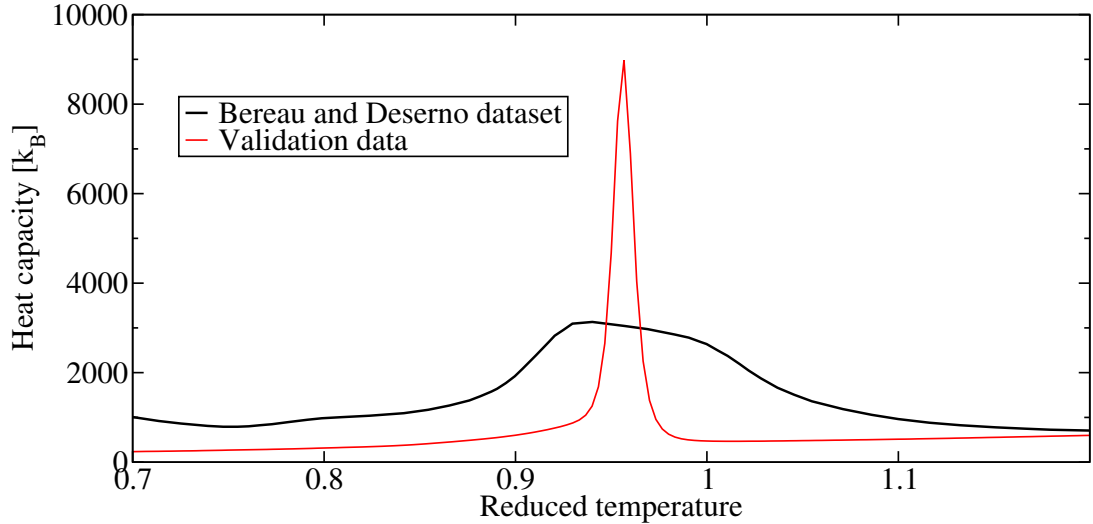


Figure 3.6: *Heat capacity of the GNNQQNY-15 system with a 40 Å periodic simulation box. Far greater sampling than the original simulation leads to the sharp peak seen in the validation dataset.*

A further study of GNNQQNY peptide systems using the PLUM model [Osborne et al., 2013] found heat capacity profiles similar to the present work, though the system sizes did not reach 15 units.

The 2A3D triple-helix bundle

The 73AA protein sequence

```
MGSWAEFKQRLAAIKTRLQALGGSEAELAAFEKEIAA
FESELQAYKKGKNPEVEALRKEAAAIRDELQAYRHN
```

is called 2A3D and is a designed three-helix bundle with a structure known from NMR [Walsh et al., 1999]. The sequence has been simulated and found to fold into the NMR native state in PLUM [Bereau and Deserno, 2009]. The system was simulated in a large box in our PLUM implementation, in a replica exchange

simulation lasting for 3.7 microseconds. The 40 thermostatted temperatures were $T_i \in \{300.0, 305.7, 311.5, 317.3, 323.3, 329.4, 335.6, 341.9, 348.4, 355.0, 361.7, 368.6, 375.7, 382.9, 390.3, 398.0, 405.8, 414.0, 422.4, 430.9, 439.7, 448.6, 457.7, 467.0, 476.4, 486.1, 496.1, 506.3, 516.7, 527.5, 538.5, 550.0, 561.8, 574.0, 586.7, 599.9, 613.8, 628.4, 643.8, 660.0\}$ K.

The system folded as expected into the native structure below the transition temperature, which was found to be approximately 398 K, based on the peak in heat capacity according to equation (1.13).

3.2.2 Statistical temperature molecular dynamics

ImpSTMD is a pair of tools for using STMD in LAMMPS. Compiling ImpSTMD produces two executables. *STMDImp_converge* runs a multicanonical simulation on a system, while dynamically updating estimates of the statistical temperature in each potential energy range, according to the STMD update scheme. *STMDImp_MC* carries out the post-convergence multicanonical simulation allowing for sampling followed by reweighting.

The implementation of STMD involves both forms of extension. The actual adjustment of forces necessary for STMD occurs within modified LAMMPS source code. LAMMPS uses the terminology of a “fix” to describe any operation involving per-time-step alteration of some property of the system, such as integrators and thermostats. The new *fix STMD* is given a pointer to a function from the driver program during its set-up. It implements a standard LAMMPS method for fixes which is called immediately after forces for the next step are calculated and communicated. This method calls the callback function to retrieve the STMD force modification factor, presented as $\frac{T_0}{T(U)}$ in equation (2.27). The method then proceeds to multiply every force in the system by this factor.

The driver itself is launched by the user and reads a configuration file detailing settings for the simulation, making use of the library *libconfig* [Lindner, 2012]. It sets up the simulation, defining the LAMMPS STMD fix, recording the pointer to the LAMMPS instance’s potential energy, equilibrating the system if specified, and so on. Once the main LAMMPS simulation begins, the driver is only returned to via the callback function, through which the driver program tracks and histograms the potential energy, characterising the system and sending back the STMD force modification factor to arrive at a uniform random walk in potential energy space.

ImpSTMD comes to approximately 2000 lines of C code and a few hundred lines of analysis tools in Python, to reweigh collective variables and produce an entropy estimate. The package also comes with a thorough user guide and examples.

Validating STMD

The STMD authors validate their work using the β -barrel BLN model [Honeycutt and Thirumalai, 1990], using the justification that “this model has been extensively studied and provides a good example of a rugged energy landscape, which cannot be correctly sampled by conventional MC or MD simulations.” [Kim et al., 2007]. The model has three categories of beads; hydrophobic (B), hydrophilic (L) and neutral (N). In particular, the 46-mer with the sequence $B_9N_3(LB)_4N_3B_9N_3(LB)_5L$, which has been extensively studied, is the validation target. This sequence folds into a globular four-stranded β -barrel at its global energy minimum [Lee and Berne, 2000].

The authors publish the system’s statistical temperature estimate, entropy estimate, and five related order parameters of the system. These parameters, labelled Q and $Q_{is,1}$ to $Q_{is,4}$, denote the structural similarity of the local energy minimum of the current configuration to the global minimum, for the whole chain, or β -barrels 1 through 4, respectively. The mathematical definition of Q is [Kim et al., 2007; Guo and Brooks, 1997; Camacho and Thirumalai, 1993]

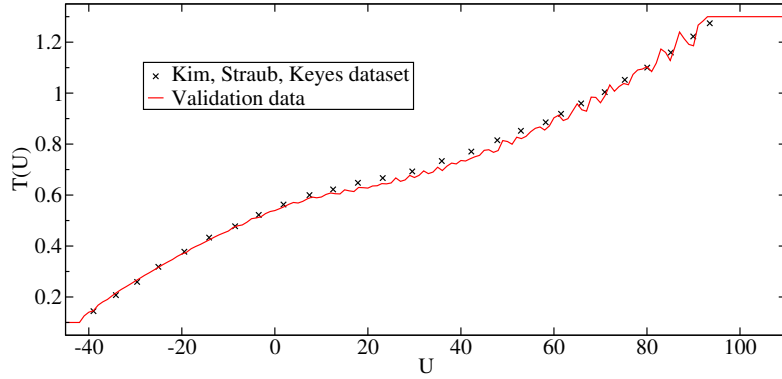
$$Q = \frac{1}{M} \sum_{i,j>i+4}^N \theta(\epsilon - |r_{ij} - r_{ij}^0|), \quad (3.1)$$

with r_{ij} and r_{ij}^0 as the relative distances between beads i and j in the current configuration and the global minimum, respectively. θ is the unit step function, with a value of 1 for a positive argument and 0 otherwise. M is a normalisation constant and ϵ is a parameter to allow for thermal fluctuations, set to 0.2 in the reference work.

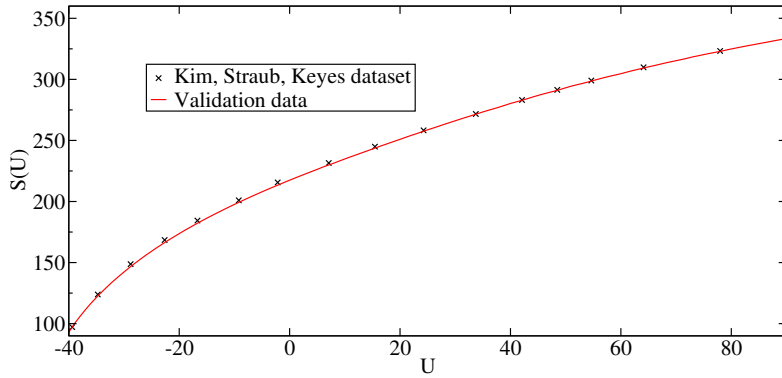
The validation data in fig. 3.7 juxtaposes the present results with the STMD authors’ results [Kim et al., 2007]. A plot digitiser has been used to recover data from image formats in the reference work. The parameter set is identical in both studies. The statistical temperature tracking array is divided into 200 bins which span the potential energy range $[-55.0, 145.0)$. The temperature range explored is $[0.1, 1.3]$, with the thermostat held at $T_0 = 1.3$. The update factor in equation (2.23) begins as $f = 1.0005$ and is reduced when the energy distribution is near-constant, having extrema within 20% of the average.

38-atom Lennard-Jones cluster

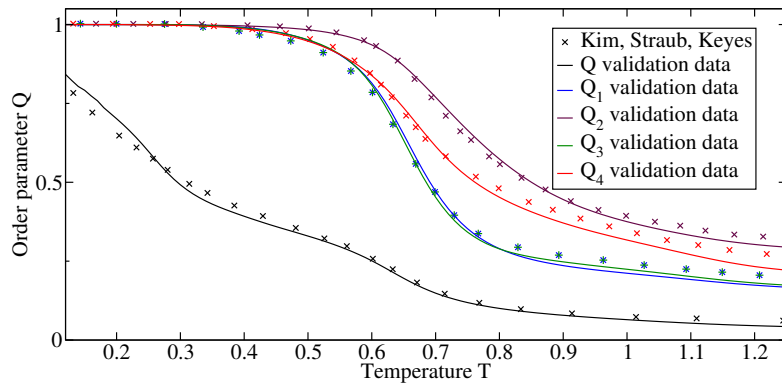
Lennard-Jones clusters are systems containing N identical particles interacting solely by the familiar Lennard-Jones potential:



(a) Convergent statistical temperature estimates $\tilde{T}(U)$. The STMD method appears to give rise to fluctuations in the estimate; note that the use of a plot digitiser has hidden a similar magnitude of fluctuations from the reference data.



(b) Entropy estimate $\tilde{S}(U)$ arising from the statistical temperature estimate in fig. 3.7a.



(c) Ensemble average estimates of a collective variable Q , describing structural similarity to the native state and defined in equation (3.1).

Figure 3.7: Validation data for the *lmpSTMD* software, comparing properties of the BLN 48-mer to equivalent work in the reference paper [Kim et al., 2007]. The original authors distinguish between staircase and linear interpolation of their running statistical temperature estimates in fig. 3.7a and 3.7b, however this does not produce a visible difference. The current work uses linear interpolation.

$$U = 4\epsilon \sum_{i < j} \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^6 \right]. \quad (3.2)$$

For ease, we operate in reduced units, setting $\epsilon = \sigma = 1.0$, $k_B = 1.0$ and referring to unitless quantities for energy, length and temperature.

A common challenge for optimisation algorithms is to locate the known global minimum of a Lennard-Jones cluster system [Pillardiy and Piela, 1995; Wales and Doye, 1997; Leary, 1997]. The 38-atom Lennard-Jones system is one of a few particularly challenging cases. There are two separate funnels leading to the lowest two minima, and the penultimate minimum is much wider and is associated with many more local minima than the global one, so relaxation from a high temperature lends itself to discovery only of the penultimate minimum [Doye et al., 1999].

A simulation was set up to test the proficiency of ImpSTMD to locate the significant minima of the Lennard-Jones 38-atom system. The global minimum is at -173.928427 [Gomez and Romero, 1994], so the window of potential energy to explore was set as $[-175.00, 13.00)$, tracked initially with 100 bins.

A variety of different starting parameters were attempted, including different update factors f in the range suggested by the STMD authors; different temperature ranges and thermostat temperatures T_0 up to 1.72, far above the core and overlayer melting point of the system [Frantsuzov and Mandelshtam, 2005]; varying bin widths and time-step sizes. In each case, the simulation became stuck in a local minimum at U_i , pushing $T(U_{i-1})$ and $T(U_{i+1})$ to the temperature caps at T_{low} and T_{high} , respectively. Example data is given in fig. 3.8.

The failure of ImpSTMD to traverse the system’s phase space suggests a problem with cool Lennard-Jones systems which could be related to the STMD scheme’s behaviour in the harmonic limit. Though it warrants further investigation, it is not necessarily an obstacle in the domain of protein modelling.

The PLUM model with STMD

The fluctuations in statistical temperature estimate seen in the validation work’s fig. 3.7a plagued attempts to utilise STMD as an accelerated sampling method for PLUM model proteins. The magnitude of the fluctuations was a high-dimensional function of the starting parameters, but two obvious parameters were seen to be most crucial; bin count and starting f_d . f_d refers to equation (2.23), simply given by $f_d = f - 1.0$.

Reducing the bin count lowers the resolution of the resultant statistical tem-

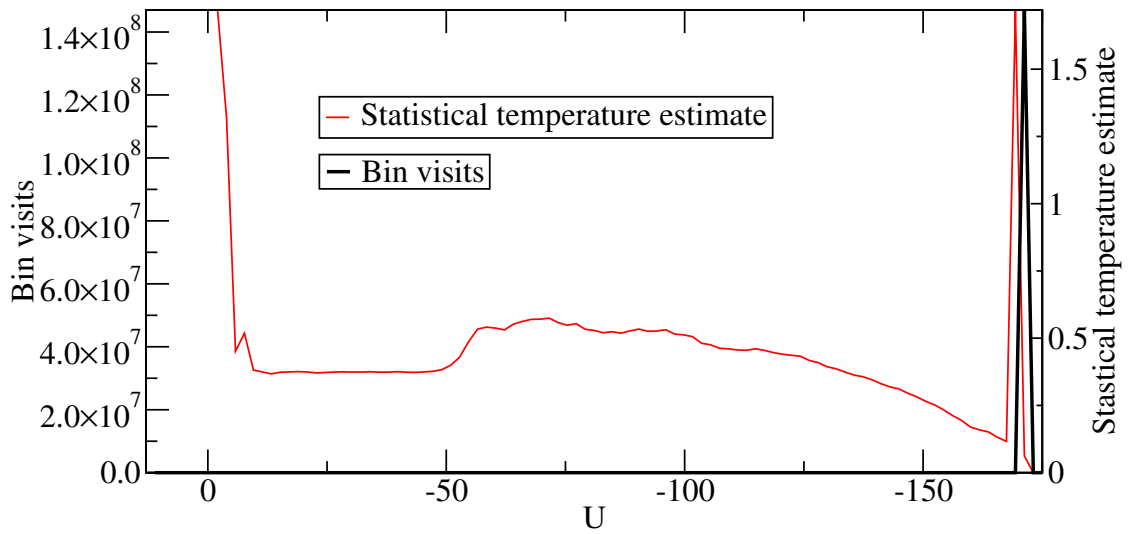


Figure 3.8: *Data snapshot from an STMD simulation of the 38-atom Lennard-Jones cluster system. Statistical temperature estimates are very far from convergence, as the simulation quickly became stuck in a local minimum configuration. This resulted in repeated erroneous revisions of the statistical temperature estimates adjacent to this bin, according to equation (2.23). The visits histogram is reset the first time a bin's statistical temperature estimate reaches the lower bound, and this is why all other bins are at zero visits.*

perature estimate, so that the smoothing it grants is inherently accompanied by a penalty in information that can be retrieved about the system. In practical terms, it means that reweighted ensemble average information about the system carries larger errors, especially in the vicinity of phase transitions.

The likely cause of the problem is that large f_d values lead to a jagged statistical temperature estimate during early exploration of the potential energy landscape, before f_d 's value is reduced to be a finer instrument. This does not hamper achieving a flat histogram, so f_d subsequently falls to a value at which it is unable to effectively fix the fluctuations within the time until the distribution is judged to be flat. Reducing the starting f_d more directly addresses the problem than lowering the bin count, but comes with a large cost in time taken for convergence, to the point that STMD may become infeasible.

The related Wang-Landau algorithm [Wang and Landau, 2001b,a] has been noted to have a similar problem, referred to as saturation in the error [Belardinelli and Pereyra, 2007a]. Modifications to its update scheme have been proposed [Belardinelli and Pereyra, 2007b] and tested with success in lattice spin models, and subsequently in lattice protein models [Swetnam and Allen, 2011].

Example data is provided in fig. 3.9, contrasting the statistical temperature estimate produced by two very different initial f_d values. The two runs have the same form, but the lower starting f_d run results in far milder fluctuations of the estimate.

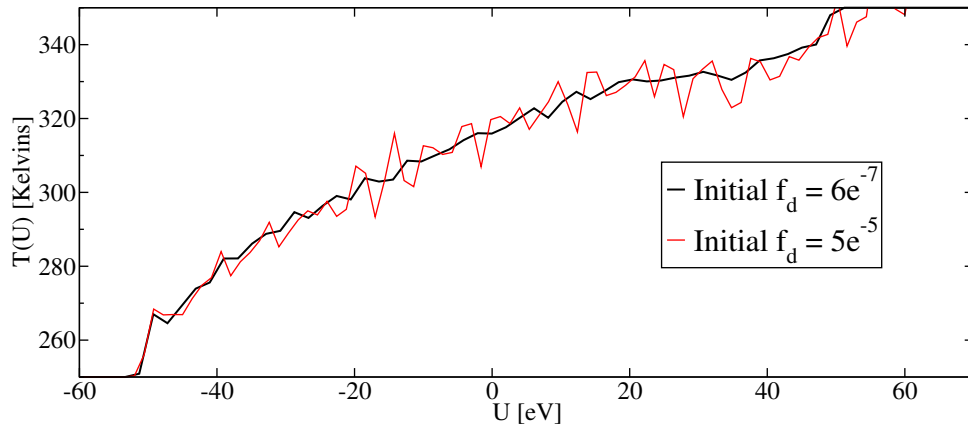


Figure 3.9: *Two STMD simulations of a single n16N peptide in a large periodic box. 100 bins span the potential energy range $[-80, 125]$, of which the sampled subsection is shown. The smoother dataset is not free of fluctuations and took 3×10^9 time-steps to reach; long enough to get a thorough sample of the system from REMD.*

Several solutions to the estimate fluctuation problem have been proposed.

Refinements of the flatness criterion and iterative f_d reduction scheme could be of use. The RESTMD algorithm [Kim et al., 2012] parallelised STMD by subdividing the temperature space with a degree of overlap, so that REMD-style swaps can occur. It may well have advantages over both of the parent methods, including making smaller starting f_d values feasible, but it was decided that implementing this scheme would be too great a deviation from the project’s core goals, with some risk of not being helpful.

The use of STMD was ultimately rejected on the basis of the preliminary studies outlined above. While workarounds to the fluctuations problem by careful choice of starting parameters seem possible, choosing an appropriate STMD parameter set itself proved to be a lengthy task involving multiple preliminary runs for any given model system. The ubiquitous REMD method is able to characterise a system in less wall-clock time than an STMD simulation, even with a carefully selected parameter set, is able to converge on a good statistical temperature estimate, which is itself a preliminary task in characterising the system.

3.3 Summary

The PLUM and PRIME20 models were implemented in the LAMMPS and DynamO packages, respectively. The PRIME20 model has many unpublished parameters; these were filled in, creating a PRIME20-like model with very similar behaviour. The accelerated-sampling method STMD had to be implemented as it is not widely available, whereas the REMD method is ready in most MD packages.

The STMD implementation in LAMMPS was tested in a range of scenarios, starting with validation and moving to PLUM model protein systems. The 38-atom Lennard-Jones system became stuck in local minima when sampled with STMD. STMD encountered problems with fluctuations in the statistical temperature estimate when dealing with a range of system types, and overcoming this obstacle was judged to be too large a time-sink to be worthwhile. Eventually, the STMD method was abandoned for the current project, in favour of the well-tested REMD approach.

Chapter 4

Single-chain simulations

It is within the capacity of explicit-water, all-atom protein modellers to sample single-chain simulations without great strain, which means that these cases can be used to check and refine the predictive power of the coarse-grained models. In this section, the simulation of several single-chain systems relevant to the project is described. Simulations are carried out in both the PRIME20-like and PLUM models, and the results are contrasted to existing atomistic simulation data. The models are adjusted as necessary to best describe the peptides of interest at this scale.

4.1 PLUM model simulations

4.1.1 Over-stabilisation of the α -helix

Simulations of a single n16N chain, denoted n16N-1, showed significant and obvious over-stabilisation of the α -helix secondary structure motif over the entire length of the chain.

A REMD simulation of the molecule was carried out with 16 replicas, each running for 8.5 microseconds. The replicas were thermostatted at $T_i \in \{275.0, 280.0, 285.0, 290.0, 300.0, 305.0, 307.5, 310.0, 312.5, 315.0, 317.5, 320.0, 325.0, 330.0, 340.0, 350.0\}$ K.

Clustering 11400 frame samples from the 300.0K trajectory into 453 groups containing geometrically alike structures (23.7% of which have a population of 5 frames or fewer) revealed that the dominant conformations are all α -helix based. Middle structures from the top four clusters are illustrated in fig. 4.1. The `g_cluster` tool available as part of the Gromacs package [Berendsen et al., 1995; Hess et al., 2008; Pronk et al., 2013], and the `gromos` clustering algorithm [Daura et al., 1999],

were used. The RMSD cut-off parameter, below which structures are considered members of the same cluster, was set to 0.4 nm.

It would be unreasonable to expect unstructured behaviour to show up as clearly as competing structured behaviour, in a structural clustering algorithm. Nonetheless, the proportion of frames taken up by α -helix structure is surprising and problematic. Ramachandran plots are provided in fig. 4.2, in order to provide an unbiased overview of the secondary structure.

The PLUM model is designed to fold into secondary structures, and here we observe that the model favours doing so even with a sequence that ought to show an absence of such structure. It may be the case that coarse-grained models are inherently too simple to capture in one parameter set both ordered and disordered protein behaviour.

The PLUM model may be more suitable for IDPs after a conservative re-tuning to less zealously seek secondary structure. The α -helix, and other common motifs, are principally stabilised by the strong energetic favourability of hydrogen bonding. This interaction will be the target for re-tuning.

In fig. 4.3, the secondary structural behaviour of the PLUM model is compared to atomistic data for the chain at different values of the backbone-backbone hydrogen bond well-depth, ϵ_{HB} . In fig. 4.3b, the data are broken down by the regions of n16N outlined in fig. 1.5 and table 1.3. Each of these datasets were obtained with the same specifications as the original n16N REMD simulation above.

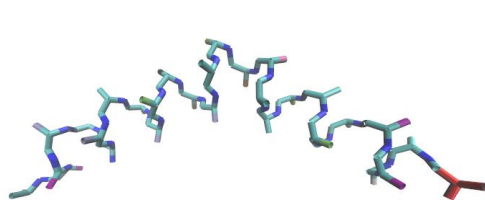
Based on this study, the decision was made to proceed using an ϵ_{HB} value set to 94.5% of the original; this will be referred to as the PLUM* model.

4.1.2 PLUM* validation work

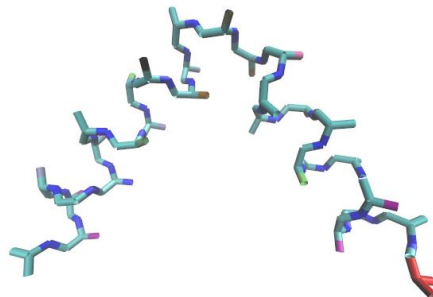
Some of the validation work on the PLUM model was repeated for the PLUM* model as a test that the model continued to function properly, without unexpected changes.

The 2A3D system in PLUM*

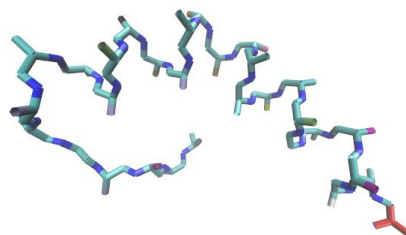
The 2A3D triple-helix bundle is simulated in PLUM* with the same conditions as its original PLUM model simulation, in section 3.2.1. The system is found to fold to the same native state, the transition to disorder now occurring at approximately 375 K.



(a) **1 (48.9%)** Residues 1 to 13 form an α -helix. A kink begins at residue 14, isoleucine, immediately preceding a highly alpha-disruptive proline. Residues 15 to 27 conform to a left-handed helix which is the enantiomer of the α -helix, maintaining $i + 4 \rightarrow i$ hydrogen bonding. The terminal residues 28 to 30 are lysine, lysine and cysteine, and these deviate from the left-handed helix dihedral values, suggesting just three residues of disorder.



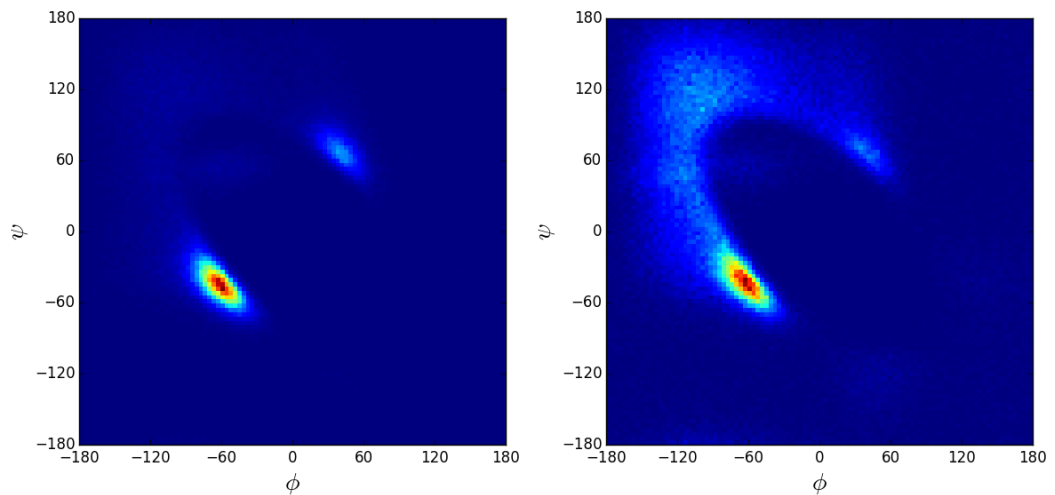
(b) **2 (10.3%)** The cluster members are α -helical throughout, with a 'double kink' centered on the proline at residue position 15. The energy penalty of the double kink is compensated for by maintenance of the favoured α structure and non-bonded interactions between the two sides of the peptide, which are pulled closer.



(c) **3 (4.1%)** A strict α -helix structure pervades the chain, with almost no variation in dihedral angles within or between structures of the cluster.

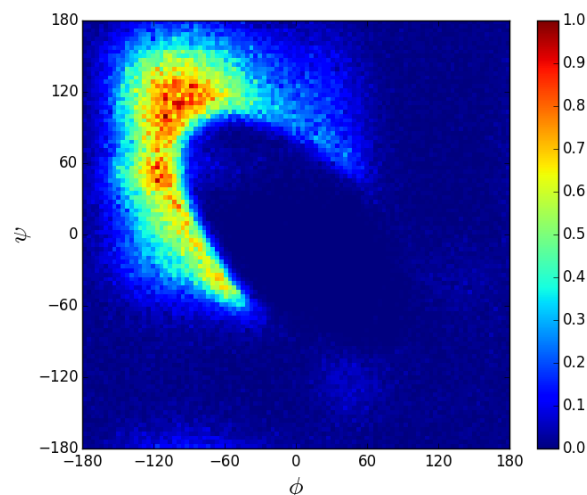
(d) **4 (3.7%)** This structure is very similar to 4.1a, having an α -helix disrupted by a kink, leading to a left-handed α -helix. In this case, the left-handed portion gives way to disorder beginning at residue 22. These C-terminal residues of the chain exhibit lower than average alpha-helicity, as made clear by fig. 4.3b.

Figure 4.1: *The four top-occurring structures for n16N represented in PLUM at 300.0K are displayed. Each structure is labelled by its rank, with the percentage of frames conforming to the structure parenthesised. The N-terminus is highlighted red and kept on the right for clarity. These top four structures cumulatively occupy 67.0% of all frames with highly ordered helical structure.*



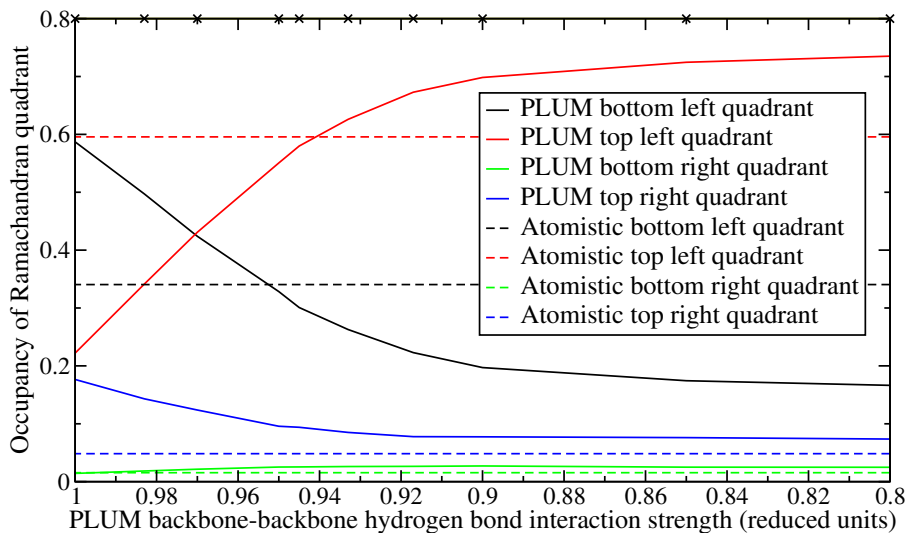
(a) **300K**: A tight, sharp global peak is seen at the coordinates for the α -helix, with a secondary peak in the left-handed α -helix domain.

(b) **325K**: The α -helix peak remains sharp and considerably dominant. The top left quadrant may have a higher overall population, however.

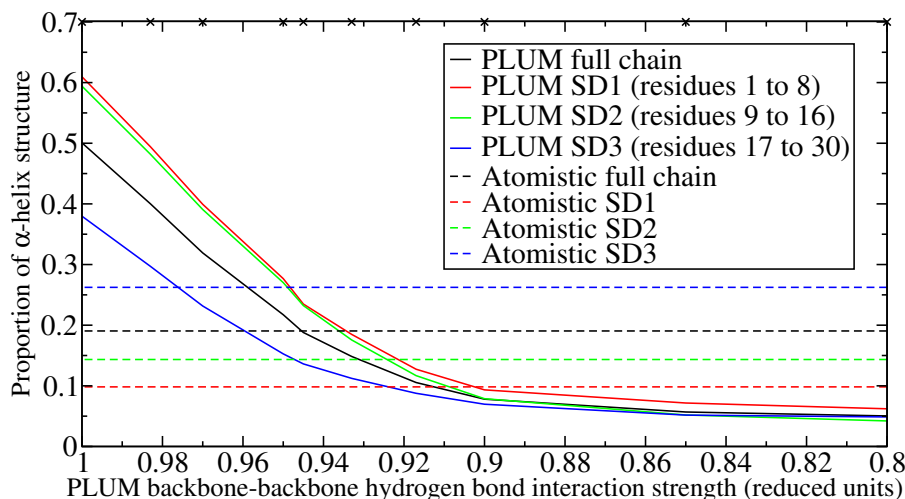


(c) **350K**: The Ramachandran plot at last becomes reminiscent of the totally unstructured GAG molecule seen in fig. 3.5.

Figure 4.2: *Ramachandran plots showing the exploration of (ϕ, ψ) phase space for a single unit of $n16N$ at various thermostatted temperatures. Even at 325K, the α -helix peak is dominant and thin. Some exploration of unfolded states and left-handed α -helix occurs at all temperatures shown, but an extremely high temperature of 350K is required for the global maximum to be in an unfolded region. This makes it clear that the PLUM model is over-stabilising this form of secondary structure for the $n16N$ molecule.*



(a) Occupancy of each of the four Ramachandran quadrants. Quadrants containing both left- and right-handed α -helices decline sharply as ϵ_{HB} falls. Most sterically permitted (ϕ, ψ) -space exists in the top left quadrant, so naturally it increases as the propensity to form α -helices falls.



(b) α -helical structure, broken down by the regions of n16N. The decline in α -helical structure follows the decline in occupancy of the bottom left quadrant shown in fig. 4.3a, and also reaches agreement with the atomistic model around a 5% decrease. However, when the level of structure is stratified by subdomain, it emerges that the PLUM model does not match the atomistic model, assigning the lowest level of α -helical structure to SD3.

Figure 4.3: Behaviour of the PLUM model of n16N at 300K as a function of hydrogen bond interaction strength ϵ_{HB} , compared to an atomistic REMD simulation [Brown et al., 2014] with the CHARMM22* model [Piana et al., 2011; MacKerell et al., 1998] in TIPS3P water [Jorgensen et al., 1983]. The PLUM output is very sensitive to adjustments, and reaches peak similarity to the atomistic data with a decrease of about 5%. Each graph has \times symbols on the top axis showing the temperatures of simulations which were run.

15-unit GNNQQNY system

The β -sheet aggregation system was simulated again in PLUM*, using the same conditions as in section 3.2.1. While the peak in heat capacity has shifted approximately from 0.96 to 0.92, the native structure below this temperature is unchanged.

4.1.3 The n16N-1 system in PLUM*

Repeating the simulation set-up of section 4.1.1, an identical clustering analysis led to the top four structures which are given in fig. 4.4. 1593 clusters of geometrically similar structures arose, compared to the previous experiment’s count of 453, and the distribution was far flatter, with top four percentages of 4.4%, 3.4%, 3.0% and 2.4% compared to 48.9%, 10.3%, 4.1% and 3.7%.

The Ramachandran plot is shown in fig. 4.5. The plot resembles the Ramachandran plot in fig. 4.2b of the canonical PLUM model at 325 K. This is unsurprising, as backbone hydrogen bonds are the key non-bonded interaction in stabilising α -helices, and scaling interaction strength is equivalent to scaling temperature. In this case, $0.945 \times 325\text{K} = 306.125\text{K}$. However, scaling a particular interaction while leaving the others untouched leads to a finer adjustment.

It is instructive to consider how well the adjusted PLUM* model now agrees with atomistic data, in aspects for which it has not been parametrically fitted. The average radius of gyration of the backbone is found to be 1.04 nm in the PLUM* model compared to 1.08 nm in CHARMM*. In fig. 4.6, secondary structure is compared to the atomistic data on a per-residue basis. The results are promising, indicating that the PLUM* model has a good ability to select between the primary options; α -structure or β -like structure, at the level of individual peptide bonds. However, the atomistic data allows a finer comparison between specific named structural regions, and this is given in fig. 4.7.

Region-wise cluster analyses were carried out on the chain for residues 1 to 8, 9 to 16 and 23 to 30 with an RMSD cut-off of 0.2 nm. The residue ranges were deliberately equal-length representations of SD1, SD2 and SD3, defined in table 1.3, making comparison possible. All region analyses produced 32 clusters. The SD1 segment’s top three clusters are populated with **41.3%**, **26.2%** and **9.5%** of frames, SD2’s with **48.1%**, **20.2%** and **9.2%**, and SD3’s with **33.3%**, **20.7%** and **15.7%**. This disparity implies that SD3 possesses the greatest conformational freedom, in agreement with the fly-casting hypothesis and with the atomistic results [Brown et al., 2014].

No regional top cluster is α -structure. SD1’s top structures are β -hairpin,

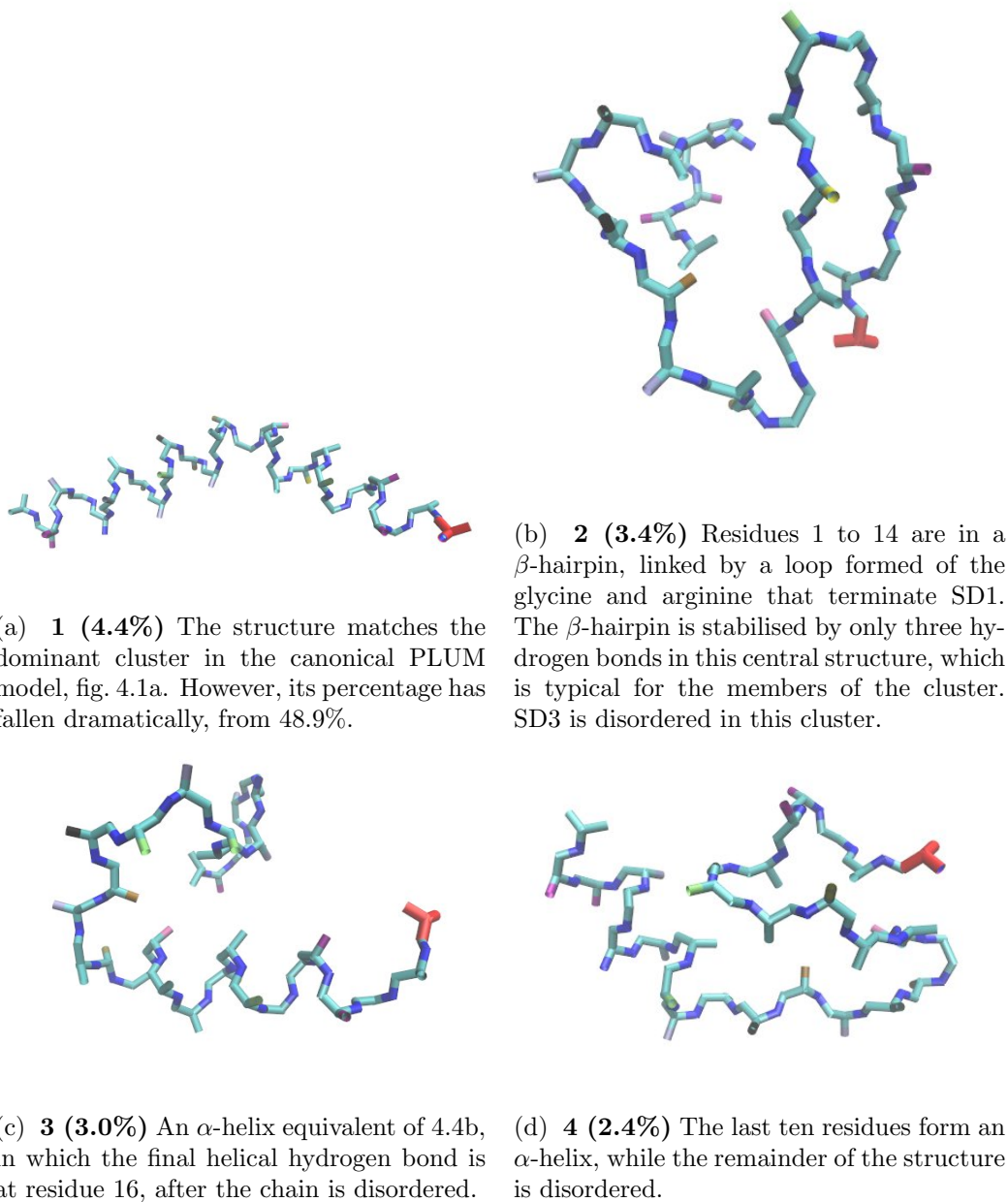


Figure 4.4: *The four top-occurring structures for $n16N$ represented in PLUM* at 300.0K are shown. The backbone hydrogen bonding parameter has been reduced to 94.5% of its canonical value. Each structure is labelled by its rank, with the percentage of frames conforming to the respective structure parenthesised. The N-terminus is highlighted red and kept on the right for clarity.*

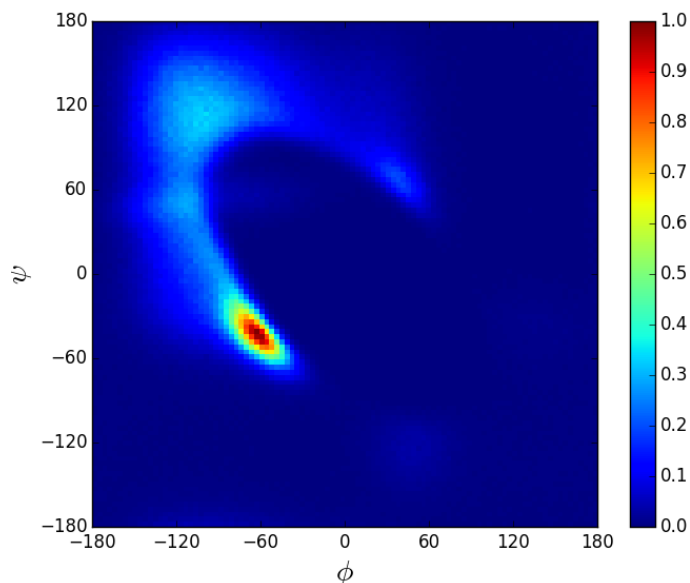


Figure 4.5: *Ramachandran plot for a single unit of n16N at 300.0K in the PLUM* model, where the backbone hydrogen bonding strength parameter ϵ_{HB} is set to 94.5% of its original value. This confirms that a higher degree of disorder now occurs, though no new peaks emerge.*

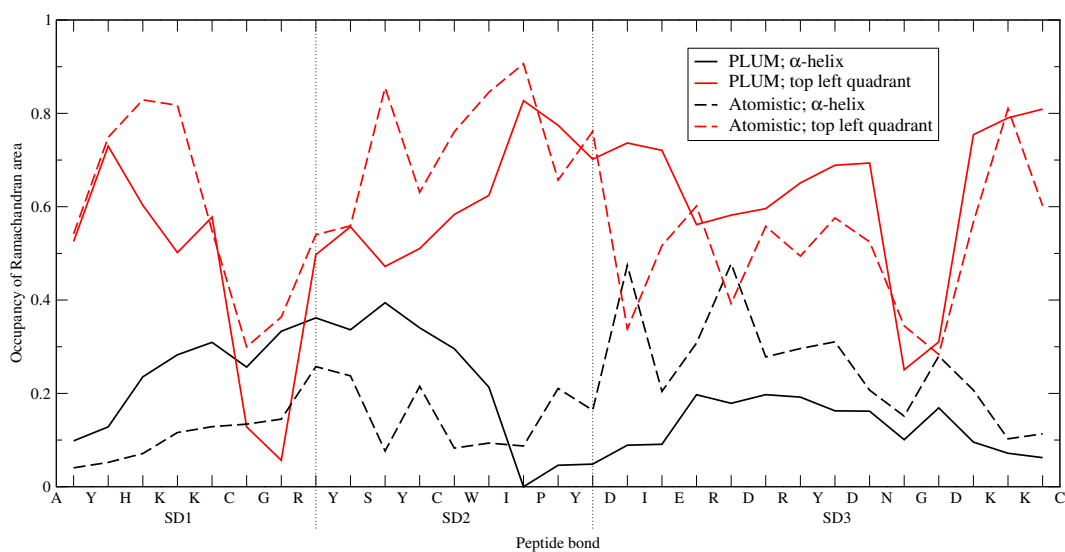


Figure 4.6: *The degree of manifestation of two different forms of secondary structure; α -helix and, broadly, “top left quadrant”, for each individual peptide bond along the chain. Proposed subdomains are demarcated by vertical dashed lines. The most striking resemblance is that each top left quadrant line is punctuated by two valleys centred on glycines, fluctuating about 0.6 otherwise. Each α line hits a minimum around SD2’s I residue, however, the disagreement in relative α -helicity between each subdomain is clear. For each pair of lines, several other minor peaks and troughs appear to line up well.*

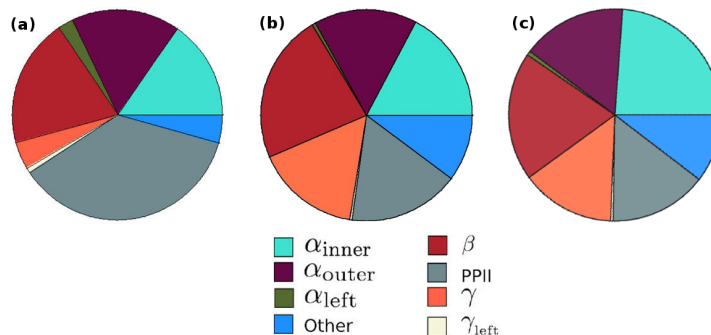


Figure 4.7: Secondary structure content of simulations of (a) *n16N* in the CHARMM22* model [Brown et al., 2014], (b) *n16N* in the PLUM* model, and (c) *n16NN* in the PLUM* model. Data represents occupancy of Ramachandran regions, without implying stable structure. The majority of segments match well in (a) and (b), but PLUM* has greater γ -structure and other structure, at the expense of PPII structure. *n16NN* has an increased propensity for α -structure, at the expense of most other structure forms. α_{left} may be under-represented in the PLUM* data, because of differences in the (ϕ, ψ) angles involved in the two models, and the authors’ choice of α_{left} region, provided in their fig. S1.

α -turn, and disorder. SD2’s top clusters are disordered, often involving a sharp turn at the residue 14I. SD3’s top clusters are all highly extended, as one might expect given its hypothesised role.

4.1.4 The *n16NN-1* system

A variant of *n16N*, with the negatively charged residues switched out for neutral substitutes (asp \rightarrow asn, glu \rightarrow gln), is known as *n16NN*. *n16NN* fails to bind with calcite and does not have substantial interactions with calcium carbonate [Metzler et al., 2008]. *n16NN* has been shown to self-assemble in an aberrant manner [Delak et al., 2007] or not at all [Metzler et al., 2010], suggesting that a simulation without ions may detect significant differences in its behaviour.

A simulation carried out in the same manner as in subsection 4.1.3 of the *n16NN* system finds subtle differences in the single-peptide conformational ensemble. A clustering analysis produces four top structures that are near-identical to those of *n16N*, with changed rankings:

1. **(8.3%)** Identical to fig. 4.4a.
2. **(3.3%)** Very similar to fig. 4.4d, with the α -helical structure persisting for an extra turn.

3. **(3.0%)** Almost identical to fig. 4.4c, except for additional α -helical structure at the start of SD3 and at the N-terminus.
4. **(2.9%)** Geometrically alike to fig. 4.4b, with no hydrogen bonds whatsoever in the “ β -hairpin” portion.

In addition to the fig. 4.7c, this points towards the conclusion that the charged residues had a role in ensuring conformational flexibility. However, region-wise cluster analyses show no clear trend towards greater local conformational freedom, suggesting that the difference lies in whole-peptide flexibility. SD1’s top clusters had populations at **40.6%**, **27.8%** and **8.1%** of frames, SD2’s at **45.9%**, **20.0%** and **15.9%**, and SD3’s at **28.1%**, **19.0%** and **18.0%**.

4.1.5 The S1 system

S1 is a bioinformatics-designed 12AA peptide, named after its place as the first peptide designed by its authors to bind strongly to quartz [Oren et al., 2007]. CD spectral analysis of the peptide shows a polyproline II structure, and this result is replicated by simulation [Oren et al., 2010]. The primary structure is

PPPWLPLYMPPWS.

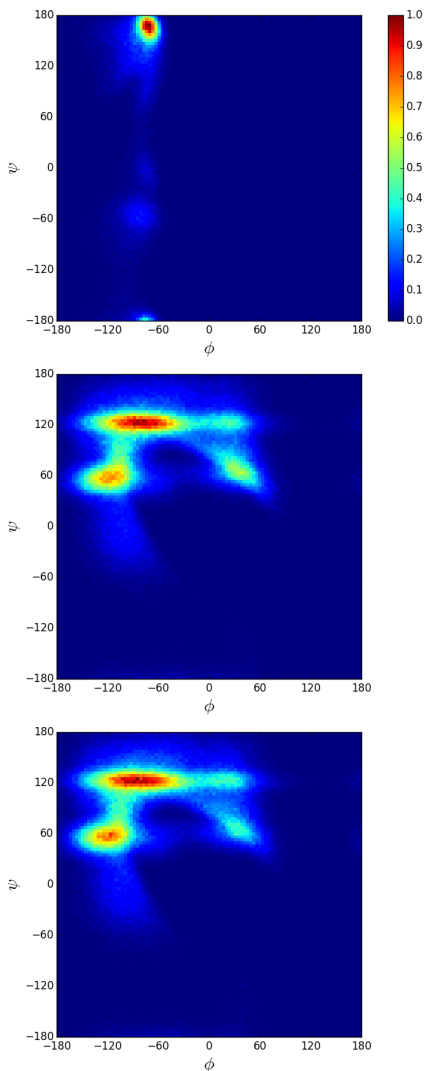
Proline, whose one-letter symbol is P, is a unique residue because its side-chain is bonded cyclically to both the $C\alpha$ and N backbone atoms, in a 5-membered ring. This makes the side-chain’s structural properties unique, limiting the ϕ dihedral angle to approximately -60° , removing the preference for *trans* isomerisation, and preventing the residue’s nitrogen from participating in hydrogen bonding.

The PLUM model features a proline residue which cannot hydrogen bond through its nitrogen and has its own ω dihedral angular potential which is bimodal, facilitating both *trans*- and *cis*-isomerisation. However, there are no further provisions, and the capacity for polyproline II structure was not specifically considered in developing the model [Bereau and Deserno, 2009].

Two sets of REMD simulations of the molecule were carried out, one in the PLUM model and one in PLUM*. 16 replicas were used, each running for 8.4 microseconds. As with n16N simulations, the replicas were thermostatted at $T_i \in \{275.0, 280.0, 285.0, 290.0, 300.0, 305.0, 307.5, 310.0, 312.5, 315.0, 317.5, 320.0, 325.0, 330.0, 340.0, 350.0\}$ K.

Fig. 4.8 shows the Ramachandran plots resulting from these simulations, in comparison to an atomistic-resolution simulation. The data reveals that the PLUM

model’s current provisions for the proline residue are insufficient to properly simulate proline-rich peptides, and further development would be required for a peptide such as S1 to be accurately characterised with the model, which falls outside of the scope of this project.



(a) Atomistic simulation data [Notman et al., 2010] conducted in the CHARMM forcefield [Brooks et al., 1983], broadly agreeing with the CD spectral data on the molecule [Oren et al., 2010]. The accessible space is limited around the $\phi = -60^\circ$ line, and the peak at $(-75^\circ, 160^\circ)$ signifies PPII structure.

(b) PLUM simulation data, showing multiple allowed regions which violate the expected structure of the chain. No peak is observed at the location of PPII structure, nor are allowed regions restricted around the $\phi = -60^\circ$ line. The major peak exists in the β -structure zone, and has a puzzling offshoot to higher values of ϕ . The third most intense peak represents α_1 structure.

(c) PLUM* simulation of the S1 peptide. The backbone hydrogen bond does not appear to be an important interaction for this peptide, as the retuning hardly affects the Ramachandran as compared to fig. 4.8b. However, the α_1 peak is weaker.

Figure 4.8: *S1* Ramachandran plots from simulation, showing a deficiency in the PLUM model for this peptide. The ideal PPII peak, present in fig. 4.8a, is accessible but not favoured in the PLUM and PLUM* simulations of this peptide.

4.2 PRIME20-like model simulations

The PRIME20 model does not establish a real-units temperature scale. In order to determine a suitable window of temperature to simulate, REMD simulations of the peptide A20 were carried out in PLUM, PLUM* and the PRIME20-like model. The PLUM simulations were carried out in LAMMPS and ran for 3 microseconds each, with 40 replicas spanning [320.0, 420.0] K. The PRIME20-like simulations ran for 1.3×10^9 events each, with 16 replicas spanning [0.1, 0.2] reduced temperature units.

In the case of PLUM, the variance of each fixed-temperature replica’s total energy E was used to deduce the system’s heat capacity at the replica’s temperature T according to equation (1.13). The DynamO package has internal tools to automatically calculate the heat capacity. The transition temperature of each system, from an α -helix structure to disorder, is known by the peak in each system’s heat capacity. These peaks fall at 386.7 K, 365.8 K and 0.168 reduced temperature units in the PLUM, PLUM* and PRIME20-like models respectively. Using these data as a rough guide, a reduced temperature of 0.13 to 0.14 is set as a plausible region for 300.0 K-like behaviour.

4.2.1 The S1 system

The S1 peptide, described in section 4.1.5, was simulated with 16 replicas spanning a reduced temperature range of [0.07, 0.22], running for 6.1×10^8 events each, in a large box. Unlike the PLUM model, the PRIME20 model has no publicly-known special provisions for the proline backbone.

The simulation was unsurprisingly unsuccessful. At low temperatures, the chain forms an α -helix, involving the prolines’ nitrogen atoms. At higher temperatures, the chain partially unfolds into a new dominant structure with some α character and some unfolded character. The transition between these states is at approximately 0.14 reduced temperature units.

This simulation work prompted us to increase the model’s realism by disabling hydrogen bonding on the proline residue’s NH bead. However, repeating the simulation with this alteration showed that it was not sufficient to lead the peptide to fold correctly. Cluster analysis was carried out on the backbone sites of the new trajectories, using an RMSD cut-off of 0.25 nm. It revealed that the chain cannot maintain an α -helix secondary motif with such significant disruptions to the hydrogen bonding. The new dominant conformation is very similar to the partially unfolded dominant conformation of the S1 simulations with proline NH hydrogen

bonding enabled, above 0.14 temperature units. It is stabilised by internal hydrogen bonding, including α -like $i + 4 \rightarrow i$ hydrogen bonding where possible, especially at the C-terminal. A family of similar partially-unfolded structural clusters remain dominant even at reduced temperature $T^* = 0.22$, at which point a phase transition appears to be underway.

The modification of disabling the proline’s NH hydrogen bonding capability is retained for the rest of the simulations and in the implementation of the model available in the DynamO package.

4.2.2 The n16N-1 system

The n16N-1 system was simulated with 30 replicas spanning the reduced temperature range [0.105, 0.250] evenly distributed with a spacing of $\Delta T^* = 0.005$. Replicas ran for at least 2.2×10^9 events each.

Within the estimated temperature window of relevance, [0.13, 0.14], and significantly beyond it, temperature appears to not greatly affect structure, and no phase transitions occur. The conformational ensembles are populated by various collapsed random coil structures with low α -helical content and relatively high anti-parallel β -strand and β -hairpin content.

The temperature-evolution of the Ramachandran plots is a gradual broadening of visited (ϕ, ψ) coordinates. Fig. 4.9 shows Ramachandran plots at three temperatures. Fig. 4.10 shows in more detail that the four quadrants of the Ramachandran plot remain fairly consistent in their populations as the temperature changes. As seen in the PLUM model, fig. 4.11 shows that secondary structure on a per-residue basis is largely congruent with the atomistic data.

Structural analysis of 10000 frames was carried out at $T^* = 0.135$, using an RMSD cut-off of 0.3 nm. Fig. 4.12 shows the top four clusters. Despite the lower cut-off, only 36 clusters arose in this analysis compared to 1593 in the PLUM* model. Even at $T^* = 0.16$, only 48 clusters arose.

A common theme of the top geometric clusters is a buried subdomain 2. The tyrosine-rich SD2 is hypothesised to have a role in both inter- and intra-peptide stabilisation, and table 4.1 bears out tyrosine’s primacy by laying out the top most prevalent residue-residue interactions, perhaps explaining why SD2 is consistently found at the core of geometric clusters.

Table 4.1 shows very high proportions of interaction between the top-interacting residues. The collapsed nature of the structures in the PRIME20-like model, which also manifests as a low number of geometric clusters being found (and a lower RMSD cut-off to distinguish them), provides a notable contrast to PLUM* and

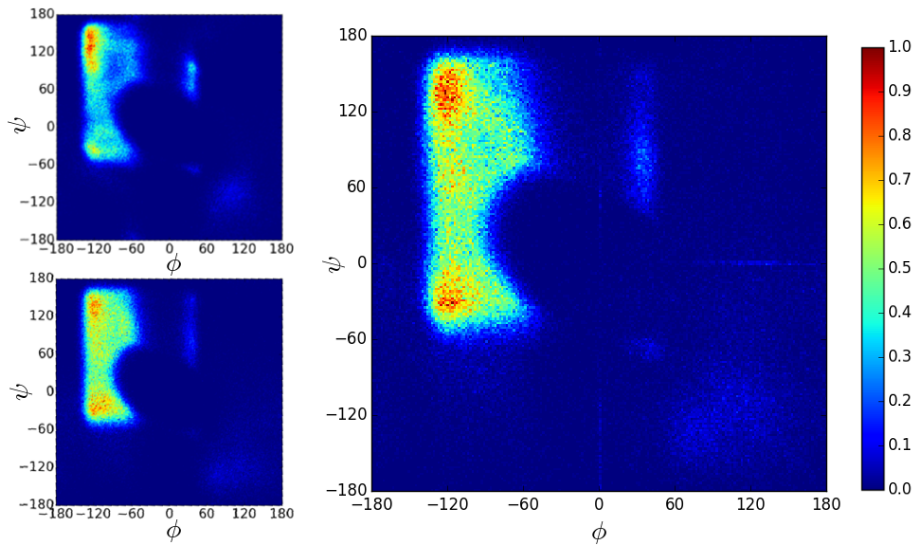


Figure 4.9: Ramachandran plot for a single unit of $n16N$ in the *PRIME20-like* model. The temperatures in reduced units are: (top left) 0.11, (bottom left) 0.16 and (right) 0.135. Two peaks are notable; one indicating anti-parallel beta structure and another at $(-121, -28)$ which does not correspond to α -structure but appears to be a turn.

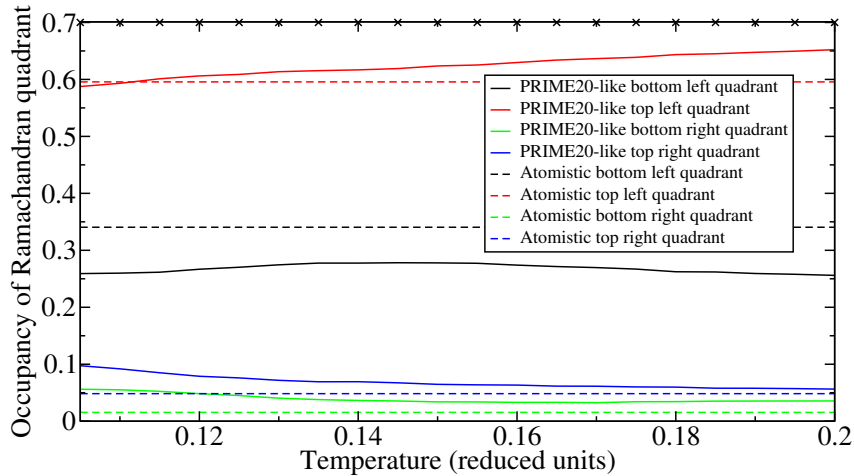


Figure 4.10: Occupancy of each of the four Ramachandran quadrants as a function of temperature. Crosses are shown on the top axis to indicate the temperatures at which data was collected. To enable comparison, atomistic data for $n16N$ at $T = 300K$ is also shown [Brown et al., 2014]. This coarse measure of structure does not show signs that the peptide changes significantly as a function of temperature. Fig. 4.9 and fig. 4.11 show that the present level of bottom left quadrant structure is disproportionately **not** α -structure.

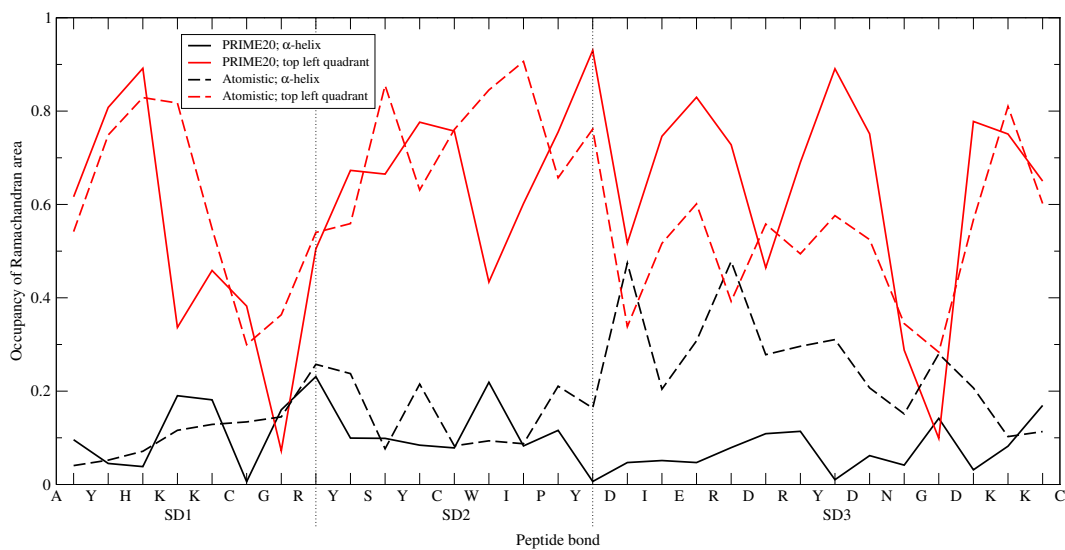
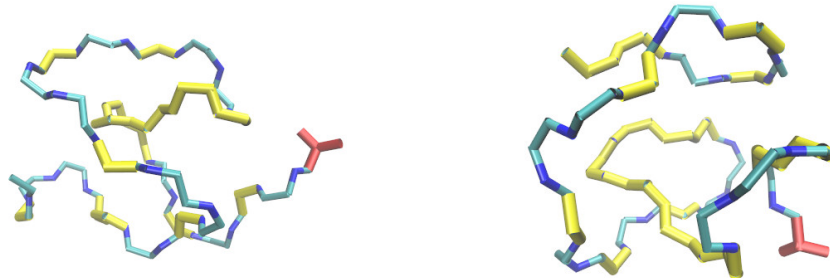
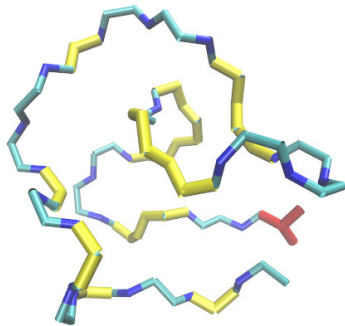


Figure 4.11: The prevalence of two different forms of secondary structure, α -helix and, broadly, “top left quadrant”, for each individual peptide bond in *n16N* simulated at $T^* = 0.135$. Proposed subdomains are demarcated by vertical dashed lines. As with the *PLUM** *n16N* data (fig. 4.6) valleys exist in the top left quadrant lines around glycine. *SD1* and *SD2* agree rather well, but, also in common with *PLUM**, a lower level of α -helical structure manifests in *SD3* in the present model than in the atomistic data.

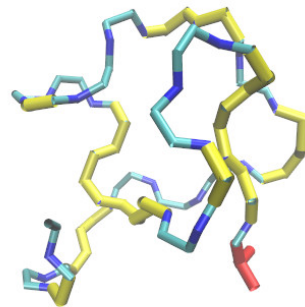


(a) **1 (11.9%)** A segment from residue 7 to residue 12 is buried in the core of the coil and achieves hydrogen bonding with outer regions on every residue, stabilising the cluster.

(b) **2 (11.1%)** Like cluster 1, a segment from residue 7 to 13 with hydrogen bonding throughout stabilises the cluster. In multiple places, the hydrogen bonds take the form of anti-parallel β -strands.



(c) **3 (10.9%)** Several hydrogen bonds bridge the gap between the core and the outer coil. Although there is no large contiguous region of hydrogen bonding in this case, SD2 is at the core of the random coil again. SD1 forms a β -hairpin structure.



(d) **4 (10.7%)** SD1 forms a β -hairpin, turning about residue K4, as in fig. 4.12c. Again, SD2 is at the core while SD3 wraps around the structure.

Figure 4.12: *The four top-occurring structures for $n16N$ represented in the PRIME20-like model at $T^* = 0.135$ are shown. Each structure is labelled by its rank, with the percentage of frames conforming to the respective structure parenthesised. For clarity, side-chains are not shown, the backbone is coloured yellow for residues involved in backbone-backbone hydrogen bonds, and the N-terminus is highlighted red and kept on the right.*

CHARMM22*. The PLUM* model is secondary structure-centric, and n16N is only weakly bound to other parts of the chain in the absence of backbone-backbone hydrogen bonds. The CHARMM22* model’s n16N has less secondary structure but is more collapsed upon itself than PLUM*’s in the absence of secondary structure. While the PLUM* model and the atomistic model produce average radius of gyration of 1.04 nm and 1.08 nm respectively, the PRIME20-like model produces an average of 0.68 nm at $T^* = 0.135$, increasing only to 0.69 nm by $T^* = 0.185$.

$T^* = 0.11$			$T^* = 0.135$			$T^* = 0.16$		
Y22	Y8	98%	Y10	Y15	87%	Y10	Y22	85%
Y15	Y8	98%	S9	Y22	87%	Y10	Y15	85%
S9	Y22	97%	Y22	H2	86%	H2	Y22	82%
Y10	Y22	94%	Y10	Y22	82%	S9	Y22	81%
Y10	Y15	93%	Y15	Y8	81%	C11	Y15	78%
C11	Y15	88%	Y22	Y8	77%	H2	S9	76%
C11	K28	87%	Y15	S9	75%	H2	Y8	75%
Y22	R7	85%	S9	C29	75%	Y22	Y8	71%
S9	C29	85%	N24	K27	75%	C5	S9	71%
H2	G6	85%	H2	G6	74%	H2	G6	69%

Table 4.1: *Top interactions between residues ranked by frequency of occurrence. To qualify, any atom on residue A has to be in the square well of any atom on residue B, with A and B separated with at least three residues in-between. Tyrosine dominates the rankings, and is clearly a cornerstone of intrapeptide stabilisation.*

As in the PLUM* model, region-wise cluster analyses were also performed on equal-length representations of SD1, SD2 and SD3. The RMSD cut-off was 0.1 nm. Regions produced 59, 30 and 61 clusters respectively. SD1’s top clusters occupy **25.4%**, **12.1%** and **11.2%** of frames, with the top and the third clusters being forms of β -hairpins, the other being an extended conformation, and all structures having a notable turn at residue K4. The β -hairpins are reminiscent of figures 4.12c and 4.12d from the PRIME20-like full-chain clustering, and figures 4.4b and perhaps 4.4d from the PLUM* clustering. A similar structure tops the list of SD1 clusters in PLUM*. SD2’s top clusters occupy **15.4%**, **14.5%** and **10.3%** of frames, and are all extended conformations stretching through the core of the coil, facilitating hydrogen bonding. SD3’s top clusters occupy **10.8%**, **10.0%** and **9.9%** of frames, and are all extended conformations which varyingly turn back to wrap around the rest of the chain. It is plausible that SD3 would stretch outwards in the presence of ions. Again, SD3 is shown to have the greatest conformational freedom of the three regions.

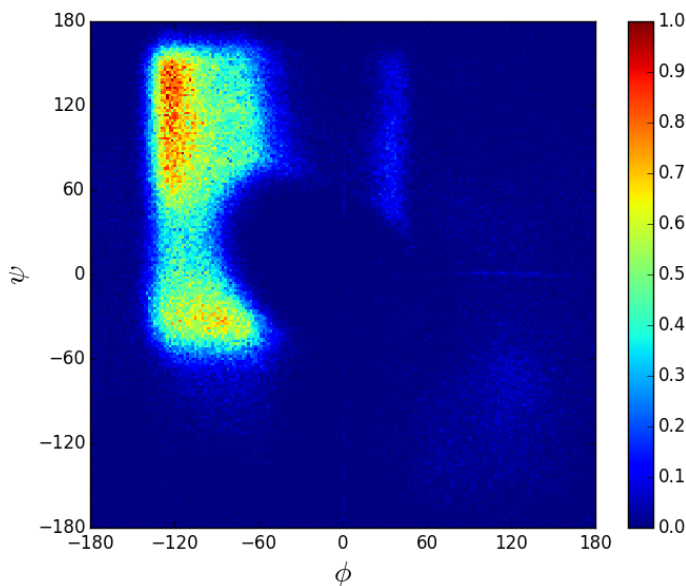


Figure 4.13: *Ramachandran plot of the n16NN peptide at $T^* = 0.135$ in the PRIME20-like model. A peak spanning a $\delta\psi$ of almost 100° is present in the β -sheet region while a much less populous peak occurs in the α -helix zone, unlike for the n16N peptide (fig. 4.9).*

4.2.3 The n16NN-1 system

The n16NN peptide, described in section 4.1.4, was simulated in the PRIME20-like model to produce a comparison to the n16N data. The simulation methodology of n16N was repeated.

A Ramachandran plot of the $T^* = 0.135$ dataset is shown in fig. 4.13. The unusual peak in the bottom left quadrant of the n16N Ramachandran data is not present for n16NN, and there is now a slight peak in the α -helix domain, which suggests the negatively charged residues were previously disrupting the potential for this structure.

The clustering analysis reinforces the notion of n16NN being more stable. In the full chain clustering, the top four clusters are populated by **23.1%**, **14.3%**, **9.2%** and **8.3%** of frames. The trend of a buried SD2 and surface SD3 from n16N is not repeated, and the top structures are irregular, wide helices with approximately 12 residues per turn, and with increasing disorder as cluster population decreases. The top structure is shown in fig. 4.14.

Top subdomain clusters have populations of **13.0%**, **8.7%** and **7.4%** for SD1, **29.2%**, **9.6%** and **5.4%** for SD2, and **19.3%**, **18.2%** and **14.4%** for SD3. Each subdomain's top structure is simply that which fits in with the 'wide helix' whole-chain structure. SD2's other structures are extended linking structures as with n16N. SD3's other structures are a β -hairpin and 1.5 turns of α -helix, terminating in disorder.

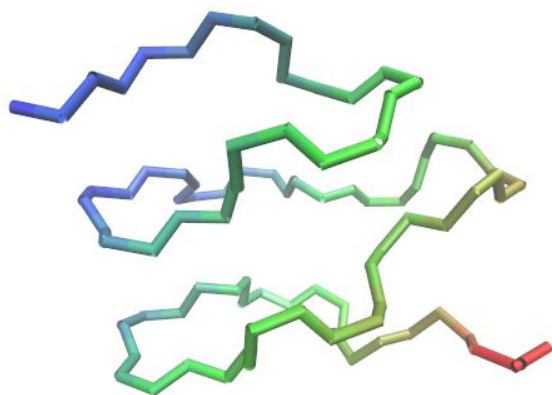


Figure 4.14: *Top structure of the cluster analysis on the n16NN peptide in the PRIME20-like model. Side-chains are not shown, and additional visual cueing is provided by the x-coordinate dependent colour scheme. The N-terminus is shown in red in the bottom right. This cluster has a frame population of 23.1% and shows an irregular helix-like structure with a large diameter.*

4.3 Summary

The PLUM model and the PRIME20-like model were used to simulate the peptides n16N, n16NN, and S1. Both models were altered to aid their accuracy for these simulations; PLUM had its backbone-backbone hydrogen bond strength reduced to 94.5% of its former value in order to alleviate the over-stabilisation of the α -helix, and the PRIME20-like model had its ability to hydrogen bond via the proline residue's backbone NH atom disabled, increasing realism of the model.

Both models proved unable to reproduce the polyproline-II helix of the designed S1 peptide, which is unsurprising as the proline residue is unique in its effect on structure, and neither coarse-grained model was developed with it in mind as a special case.

Both models had some success with ensemble average structural properties of the n16N system. Measures of secondary structure based on dihedral angles showed a good approximation to atomistic data, both on average for the chain and on a per-peptide bond basis. Measurements of top residue-residue interactions showed the importance of tyrosine-tyrosine interactions was replicated in the PRIME20-like model.

Geometric clustering produced results in agreement with existing data, showing SD3 to be the most disordered region, and showing full-chain conformational flexibility to fall as expected in the mutant n16NN. However, PLUM*'s top structures remained anomalously α -helix based, while the PRIME20-like model had a conformational ensemble of structures that were more collapsed than the atomistic

data.

Chapter 5

Multiple-chain simulations

Due to limits on computational power, little work exists characterising the behaviour of larger protein systems. This is the domain in which coarse-grained models become relevant. In this section, simulation data of n16N and n16NN in multiplicity are presented. The chapter structure changes to system-first (e.g. n16N-2) rather than model-first, as the focus is now on the systems and not the models.

We continue to use the `g_cluster/gromos` clustering tool in this work, with one custom improvement: The program now accepts an integer argument giving the number of identical molecules the system is made of, and a trajectory frame can be added to a cluster if any permutation of the chains' positions within a frame lead to an RMSD below the cut-off.

5.1 The n16N-2 system

Two units of the n16N peptide were simulated in both the PRIME20-like and PLUM* models. This system is called n16N-2. As n16N is known to form macromolecular complexes which may make use of specific aggregation domains, may contain different secondary structure from the monomers, and whose formation may be essential for n16N's biomineralisation properties, the structural properties of the aggregates is an area of great interest.

5.1.1 PLUM*

A REMD simulation of the n16N-2 system was carried out with 30 replicas, each running for 5.1 microseconds. The replicas were thermostatted at $T_i \in \{275.0, 278.54, 281.69, 284.57, 287.23, 289.67, 291.84, 293.8, 295.61, 297.29, 298.87, 300.00, 301.79, 303.15, 304.46, 305.73, 307.06, 308.47, 309.98, 311.61, 313.39, 315.39, 317.69,$

320.5, 324.19, 328.27, 332.72, 337.66, 343.3, 350.0} K.

34400 trajectory snapshots were used for geometric clustering at 300.00 K, which was carried out with an RMSD cut-off of 0.6 nm. 11992 clusters were found, the top of which had populations of 2.6%, 1.9%, 1.0% and 1.0%, and these are shown in fig. 5.1. The analysis shows a trend of subdomains SD1 and SD2 being kept in rigid conformations at the core of the dipeptide system, holding the chains together, while SD3 is extended, uninvolved in interpeptide interactions, and significantly more able to sample different conformations. The region-wise cluster analysis makes this even clearer. An RMSD cut off of 0.2 nm is used, and chains are examined one at a time (i.e. each frame of two peptides is decomposed into two frames of one peptide). SD1's top clusters' populations are **55.4%**, **10.6%** and **8.0%**. SD2's are **68.4%**, **8.7%** and **8.2%**. SD3's are **37.5%**, **20.0%** and **18.0%**. All subdomains have 35 geometric clusters in total.

Fig. 5.2a shows the proportion of frames for which each residue is involved in interpeptide interactions. Combined with the clustering analysis, these data are in striking agreement with the hypothesised domain roles; table 1.3. SD1 and SD2 are highly involved in interpeptide stabilisation, with the tyrosine-rich SD2 being most involved, both by backbone and by side-chain interactions. Interpeptide interactions decline after SD2, so that the tail of SD3 is largely free and unbound.

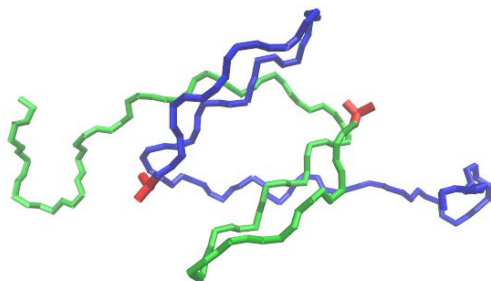
Fig. 5.3 shows the Ramachandran plot of all (ϕ, ψ) angle pairs in the n16N-2 system, as well as a second plot, highlighting the difference in folding behaviour between n16N-1 and n16N-2. These data show that α -helix structure is far less favoured in the n16N-2 system than in n16N in isolation, and β -structures now form the greatest peak.

The fact that multiplicity of the peptide in the system vastly changes the peptides' folding and draws divergent behaviour out of each subdomain, aligning with hypothesised aggregation-dependent function and featuring disorder, is a remarkable property of the system for Bereau and Deserno's four-bead model to capture. This result suggests a strong possibility of a role for coarse-grained protein models of this level in studying intrinsically disordered protein behaviour.

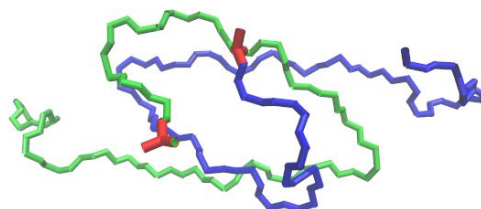
5.1.2 PLUM

Prior to finalising the decision to retune an aspect of the PLUM model to improve modelling of IDPs, data was collected on the n16N-2 system in PLUM. The same simulation parameters were used as in the PLUM* simulation, above. To facilitate comparison, the PLUM data will be briefly summarised and discussed here.

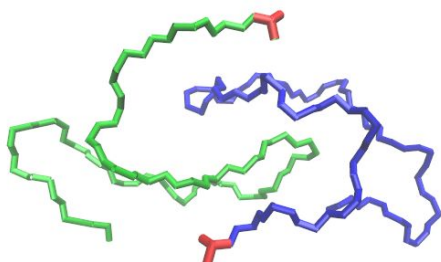
A clustering analysis carried out equivalently to that of n16N-2 PLUM*



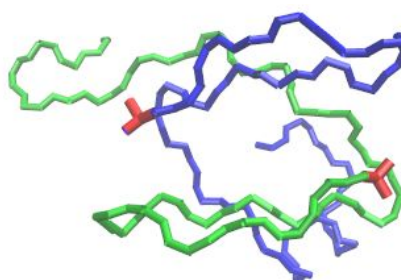
(a) **1 (2.6%)** Both peptides' SD1 and SD2 form β -hairpins, turning on residues G7 and R8, and ending with a turn on I14 and P15. Each chain passes under the other's hairpin and competes for hydrogen bonds. Several highly favourable interpeptide side-chain interactions are in register, primarily in SD1 and SD2, but also involving SD3's I and Y residues.



(b) **2 (1.9%)** Parallel β -strands bind the chains, with few intrapeptide interactions. Turns occur messily around residues K5 to R8 and more sharply at residues I14 and P15. Interpeptide hydrogen bonding and side-chain interaction continues as far as Y23, so that, as in fig. 4.12a, only the final seven residues form a free tail.

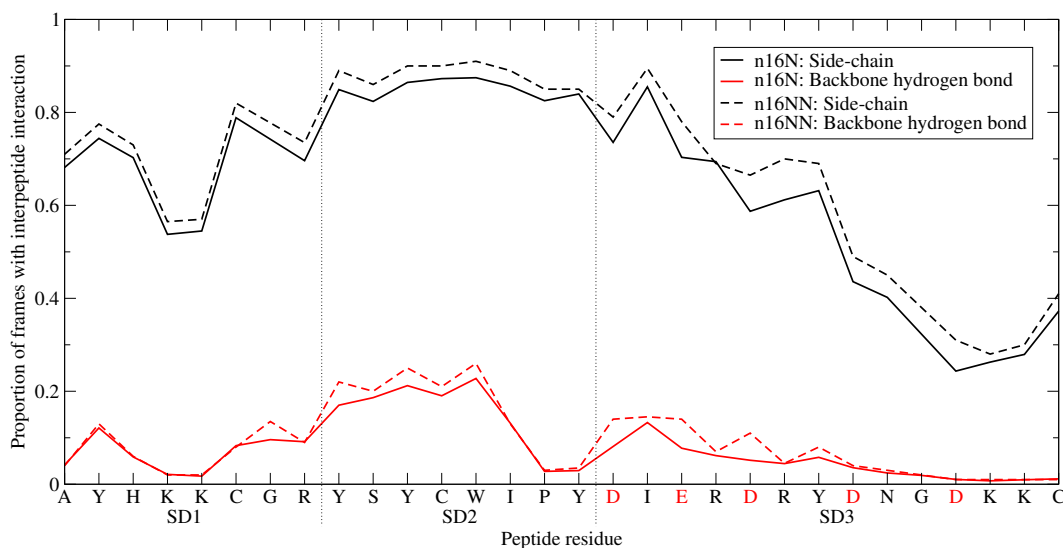


(c) **3 (1.0%)** The five residues Y2, S10, Y11, C12, and W13 form interpeptide hydrogen bonds in a straight line from one chain's Y2 to the other's, and this is the only interpeptide hydrogen bonding. A great deal of side-chain interactions stabilise the two peptides, and, again, Y23 is the last residue involved in these.

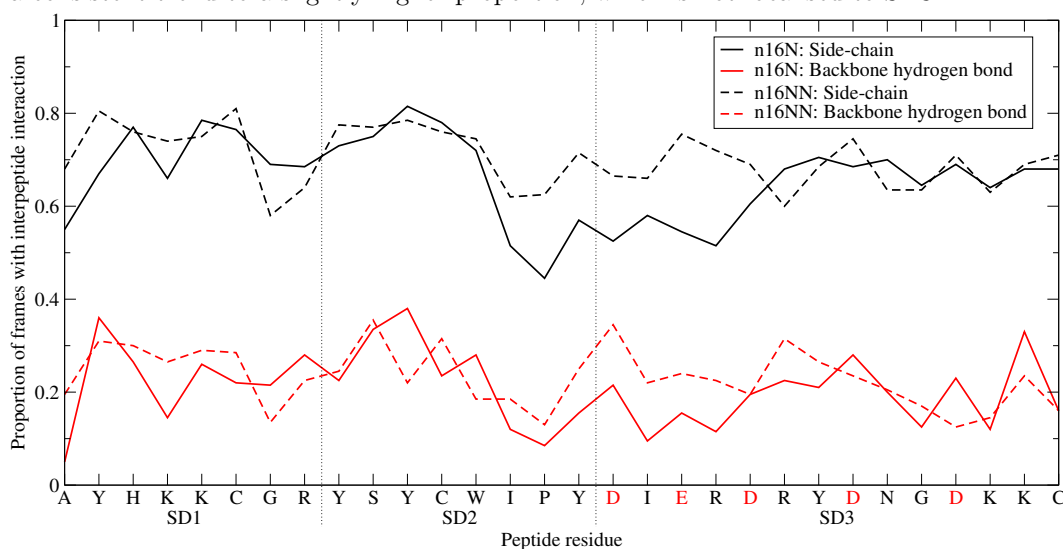


(d) **4 (1.0%)** Extremely similar structure to fig. 5.1a, with slightly different tail structure. The blue chain's tail forms two turns of an α -helix from residues E19 to G26, while the green ends in a β -hairpin.

Figure 5.1: *The four top-occurring structures for the n16N-2 system in PLUM* at 300.0K. Each structure is labelled by its rank, with the percentage population given in brackets. N-termini are highlighted in red. In all structures, the N-terminal half is central to the structure while the C-terminal half forms a tail. Differences in side-chain interactions must be responsible for this asymmetry.*



(a) **PLUM***: In agreement with observations from geometric clustering analysis (fig. 5.1), the first two domains are far more involved in interpeptide stabilisation than the third. The proportion reaches a maximum in SD2. Residues with the most favourable hydrophobic interactions; I, W, C and Y, and their neighbours, have higher proportions. n16NN shows a consistent trend to a slightly higher proportion, which is not localised to SD3.



(b) **PRIME20-like**: Each line fluctuates about a fairly flat moving average, and the only noteworthy difference is a valley passing through the SD2/SD3 boundary in both hydrogen bond and side-chain lines and lasting approximately 5 residues into SD3. The mutations from n16N to n16NN cause one large effect; the disappearance of these valleys. Otherwise, the structural ensemble alters enough to cause each value to shift pseudo-randomly, but has the overall effect of making each line even flatter.

Figure 5.2: *The proportion of trajectory snapshots for which any given residue along the chain is involved in an interaction binding it to the other chain, in the n16N-2 and n16NN-2 systems. The n16N residue sequence is shown on the x-axis; the red residues are replaced in n16NN according to $D \rightarrow N$ and $E \rightarrow Q$. Interactions are divided into side-chain and hydrogen bond types; an interaction for glycine is not always applicable, and in these cases, no data-point is plotted. Note that the disparate forms of interaction in each model make a comparison of the average of each line meaningless.*

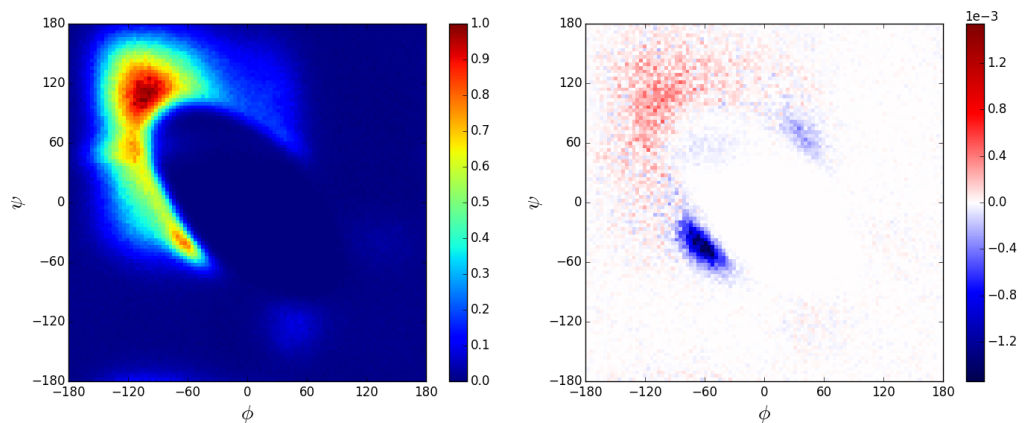


Figure 5.3: *Ramachandran plots of the n16N-2 system. **Left** shows a standard Ramachandran heat map. The greatest and widest peak is for β -strand structure, while a minuscule peak exists for α -helix structure, with a strong pathway between the two. **Right** shows a difference heat map, with the Ramachandran heat map of PLUM* n16N-1 (fig 4.5) normalised appropriately and subtracted from the heat map plotted on the left. In the scale shown, a value of -1.0 would imply 100% of hits being in a given bin in the n16N-1 simulation, and 0% in the n16N-2 simulation. Two absences are revealed in the locations for α and α_{left} structure, while β -structure is far more prominent. This implies that β -structures involving interpeptide interactions are more stable than intrinsically intrapeptide α -helices.*

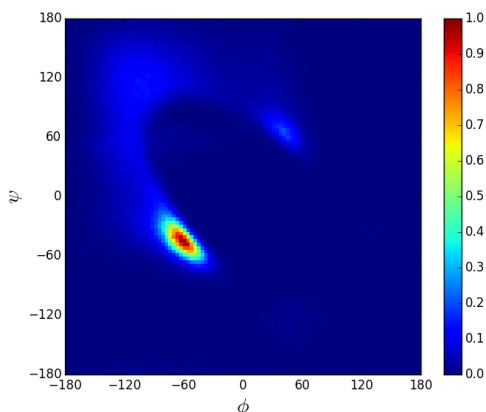


Figure 5.4: *Ramachandran heat map for the n16N-2 system in the unaltered PLUM model. Unlike in the PLUM* model, n16N in PLUM does not shift away from α -helix structure in the dimer system.*

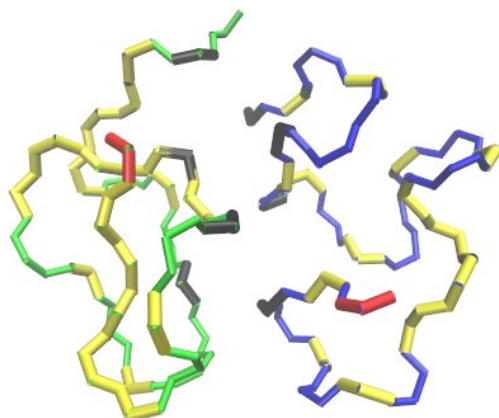
shows that the top four clusters have populations of 3.2%, 2.3%, 2.0% and 1.8%, and all of these but the third are two α -helix structures as in n16N-1 PLUM (fig. 4.1), associated with each other in different configurations via side-chain interactions, while almost all backbone interactions are intrapeptide. The third structure matches fig. 5.1a closely. Differentiation of the subdomains therefore occurs to a far milder degree, and aggregation is mostly limited to side-chain interactions supporting existing secondary structure. The Ramachandran plot of fig. 5.4 shows that the α -helix peak remains completely dominant.

5.1.3 PRIME20-like

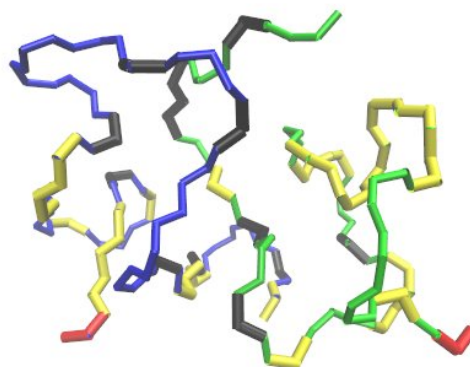
30 replicas were used to simulate the n16N-2 system in DynamO, in a REMD set-up spanning the reduced temperature range [0.105, 0.250] evenly distributed with a spacing of $\Delta T^* = 0.005$. Replicas ran for at least 9.0×10^9 events each.

18788 frame snapshots were saved and used for geometric clustering, at $T^* = 0.135$, with a RMSD cut-off of 0.5 nm. The top four clusters are shown in fig. 5.5. Unlike the PLUM* results, these data do not conform to the subdomain hypothesis laid out in table 1.3. Other than fig. 5.5c, the top structures appear to be much like single-peptide structures, merely perturbed by sitting next to each other. Like the PLUM* data, no frames exist in which the two chains are uninvolved with each other by any interaction.

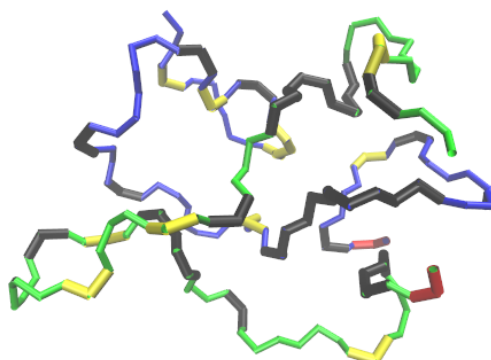
Fig. 5.2b shows the proportion of frames in which each residue is involved in interpeptide interactions. This figure gives a more complete overview of how each chain interacts with the other. Any preference for interacting via a particular subdomain is harder to see than in the PLUM* data, backing up what the top clusters show. The global maximum for both n16N lines does correspond to the Y11 residue in SD2, very similar to the PLUM dataset.



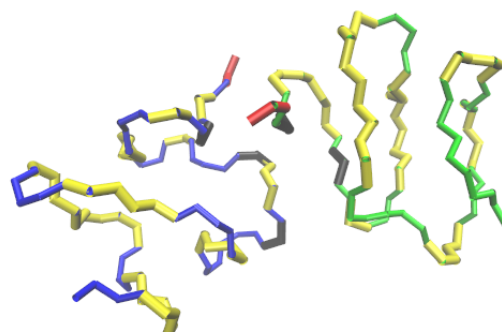
(a) **1 (8.2%)** Each peptide forms a random coil, interacting with the other through a planar interface. The structure is asymmetrical, and the vast majority of interactions are intrapeptide. SD2 regions are the least involved at the interface, while each chain's SD1 interacts predominantly with the other's SD1, and the same is true for SD3.



(b) **2 (7.9%)** Chains' SD3s are heavily involved in interpeptide interactions, and are at the centre of the chain system. This is a reversal of the situation seen in the PLUM* model and predicted in the subdomain hypothesis. Each chain remains a random coil structure.



(c) **3 (7.4%)** The chains are highly involved with each other in an asymmetrical manner, and with no apparent subdomain preference for interpeptide interaction. However, each chain's subdomain interacts primarily with the same subdomain on the other chain.



(d) **4 (6.9%)** Each chain adopts a 'wide helix' structure, previously seen in n16NN-1 in the PRIME20-like model, fig. 4.14. Only three interpeptide hydrogen bonds exist, making this the top structure with the fewest interpeptide interactions.

Figure 5.5: *The four top-occurring structures for the n16N-2 system in PRIME20-like model at $T^* = 0.135$. Each structure is labelled by its rank, and the percentage population is given in brackets. N-termini are highlighted in red. Interpeptide and intrapeptide hydrogen bonding are highlighted in black and yellow respectively.*

A subdomain geometric clustering analysis was also carried out, using the usual method of representing SD1, SD2 and SD3 as equal-length chain segments; 1 to 8, 9 to 15 and 23 to 30. An RMSD cut-off of 0.1 nm is used. Unlike all previous n16N simulations, SD3 here actually shows the least flexibility, its top clusters accounting for **10.2%**, **7.4%** and **5.6%** of frames, compared to SD1 at **5.1%**, **4.8%** and **4.5%**; and SD2 at **8.7%**, **7.1%** and **6.2%**. The previous trend of SD1 being populated by extended and β -hairpin structures, turning about residue K4, continued. SD2's top clusters involved single α -turns in the top two clusters, and are otherwise extended. SD3's top cluster is a β -hairpin, and the next two are disordered.

The Ramachandran plot for the system is shown in fig. 5.6, along with a difference heat map showing how secondary structure has changed in the dimer system compared to the monomer. The absence of significant differences that the different plot highlights underlines the nature of aggregation of these chains in the PRIME20-like model, which is first hinted at by the cluster analysis: The chains primarily interact with themselves, and the presence of the other chain makes no drastic changes to structure.

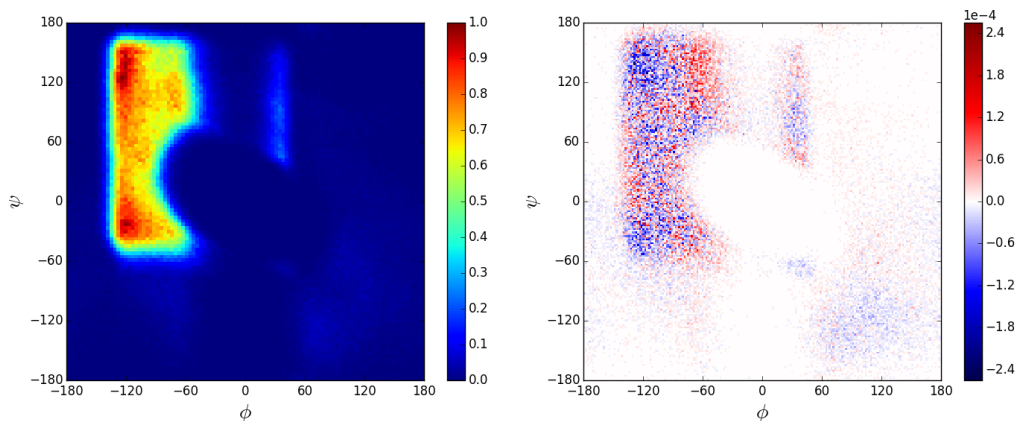


Figure 5.6: *Ramachandran plots of the n16N-2 system. **Left** shows a standard Ramachandran heat map. As with n16N-1 in PRIME20-like (fig. 4.9), there are two notable peaks, one indicating anti-parallel β -structure, and the other which has been characterised as a turn. Compared to n16N-1, the peaks have broadened greatly. **Right** shows a difference heat map, highlighting the difference between the n16N-1 and n16N-2 systems in (ϕ, ψ) angles adopted. While it appears that some systematic changes have occurred, these are much harder to distinguish from noise than the PLUM* case (fig. 5.3), and may simply be peak broadening. The differences also have about a fifth the magnitude of the PLUM* case.*

5.2 The n16NN-2 system

As it has been reported that n16NN aggregates in an aberrant manner [Delak et al., 2007] or not at all [Metzler et al., 2010], a simulation of two units of n16NN, known as the n16NN-2 system, may be sufficient to find significant differences in behaviour from n16N.

In section 4.1.4, slight differences in the preferred structure of n16NN compared to n16N in the PLUM* model were demonstrated, including a minor rearrangement of the top favoured geometrical clusters, and a lower degree of whole-chain conformational flexibility. Section 4.2.3 shows greater distinctions between n16NN and n16N in the PRIME20-like model. There is a total upheaval of the favoured structures, and, again, full-chain flexibility is lower in n16NN.

5.2.1 PLUM*

The simulation parameters employed for the n16N-2 system in section 5.1.1 were repeated here, the replicas now running for 6.4 microseconds. As with the single-chain systems in PLUM*, the differences in structure are present, but minor.

A geometric clustering analysis of the whole chain, using an RMSD cut-off of 0.6 nm, finds that the top most populated clusters are all extremely similar to n16N-2 PLUM* structures, except for differences in SD3, as follows:

1. **(6.9%)** Fig. 5.1b.
2. **(2.7%)** Fig. 5.1a, with SD3s collapsed against the central structure.
3. **(0.94%)** Fig. 5.1a.
4. **(0.82%)** Fig. 5.1a, with α -helix tails.

The n16N-2 and n16NN-2 systems were found to be the sums of a stable core, made of the N-terminal halves of the chains, and unstable flailing tails made of the C-terminal ends. The change from n16N to n16NN involves point mutations in the C-terminal half of the chain, not directly affecting the properties of SD1 and SD2, and yet the top core structures show up in different order with different weights in the full-system clustering analysis. This could be a skew resulting from the changes in SD3, or it could be inherent. Therefore, the decision was made to geometrically cluster the atoms typically corresponding to the stable core, independent of the skewing effects of the tails. Specifically, the regions comprising SD1 and SD2 in the region-wise analyses were joined and an RMSD cut-off of 0.3 nm was used. Data from both n16N and n16NN were studied in this condition.

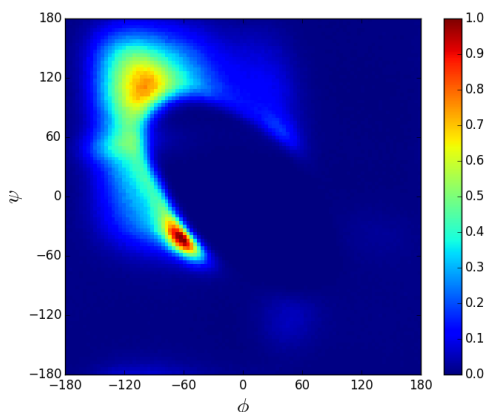


Figure 5.7: *Ramachandran heat map for the n16NN-2 system in the PLUM* model at 300 K. The global maximum exists in a sharp peak for α -helix structure, and a broader peak exists for β -structure.*

For both n16NN-2 and n16N-2 systems, clustering the cores alone resulted in the top most popular clusters matching the cores of n16N-2, clustered as a full system (fig. 5.1). However, the n16NN-2 system showed a much greater degree of core stability, its top clusters scoring populations of 3.5%, 3.2% and 1.3%, compared to 2.6%, 0.79% and 0.63%. This complements the result seen for single units of the peptides that full-chain flexibility, not just SD3 flexibility, drops as a result of the changes from n16N to n16NN.

The proportion of frames for which each residue along the chain is involved in interpeptide hydrogen bonds is plotted in fig. 5.2a. A surprisingly simple difference is seen between the n16N and n16NN lines, which is a slight increase in proportion throughout, once again lending strength to the hypothesis of the SD3 changes having a full-system effect.

As with the single-peptide systems, no clear differences in regional peptide flexibility were seen between n16N-2 and n16NN-2 via the regional clustering analysis, carried out in the same manner as in the n16N-2 PLUM* system.

A Ramachandran plot of the n16NN-2 system in PLUM* is included in fig. 5.7. Despite the increased stability of the α -helix-free core structures demonstrated above, the Ramachandran plot shows a relatively strong degree of α -helicity. This was found to stem from the C-terminal half of the chain, which is far more likely to manifest α -motifs in the absence of the negatively charged residues.

5.2.2 PRIME20-like

The n16NN-2 system was simulated in a large box for 6.6×10^9 events. 30 replicas ran in a REMD set-up, spanning the reduced temperature range [0.105, 0.250] evenly.

The Ramachandran plot for this system is in fig. 5.8. The plot shows far broader peaks than the n16N-2 system; evidence that the present system is more

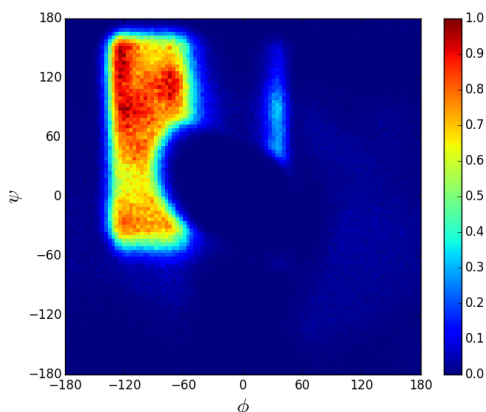


Figure 5.8: *Ramachandran heat map for the n16NN-2 system in the PLUM* model at 300 K. The largest island of allowed (ϕ, ψ) values is extremely flat, having no strongly preferred regions, though its maximum is in the location of anti-parallel β -structure, as with n16N-2.*

strongly random coil in character. Further evidence for increased homogeneity is provided by fig. 5.2b, which shows the rate of interpeptide interaction by residue. The only clear feature in the line for n16N-2 is a wide valley passing through the boundary between SD2 and SD3; this disappears for n16NN-2, and the line becomes scarcely distinguishable from noise.

The full-chain cluster analysis was carried out with an RMSD cut-off of 0.5 nm again, and continued the trend of n16NN showing a slightly lower level of flexibility than n16N. The top clusters score 10.1%, 8.3%, 7.5% and 6.9% and are composed, in the main, of two random coils interacting via an interface.

Subdomain clustering analyses were carried out according to the usual protocol, with the RMSD cut-off set to 0.1 nm. Compared to n16N-2, a significant shift towards disordered and extended conformations was observed, agreeing with the interpretation of the Ramachandran plot presented here. No subdomain has a recognisable secondary structural motif as the most popular structure, and only SD2 features any in the top three structures, its second cluster having a turn of α -helix.

The top three clusters of each subdomain have populations of (SD1) 5.0%, 4.0%, 3.9%; (SD2) 9.5%, 6.7% 5.2%; and (SD3) 7.6%, 6.2%, 4.7%. This shows no clear trend in local chain flexibility, but, surprisingly and uniquely, SD3 appears more flexible in n16NN in this case.

5.3 Discussion of dimer systems

The PLUM* model and the PRIME20-like model were used to simulate the systems n16N-2 and n16NN-2, which denote two units of the peptides n16N and n16NN, respectively.

The PLUM* model exhibited an ability to distinguish between the local primary structure of different parts of the chain and cause different behaviour to manifest accordingly. Residue-by-residue analyses and regional analyses showed locally specific structure and function, correlating with hypothesised domain-dependent roles. [Brown et al., 2014]’s three-domain scheme for the peptide, delineated in table 1.3, was bolstered by evidence from PLUM*, especially with respect to the distinction between SD1 and SD2 as aggregation-enablers and SD3 as a free tail. These differences between the subdomains were hidden in the single-peptide systems and required peptide multiplicity to come to light.

Compared to n16N-2, the n16NN-2 system featured a far more stable SD1 and SD2 core; an interesting result, as these subdomains are identical in the two peptides. As it has been suggested that disorder is useful for molecular assembly (see sec. 1.2.2), this may be relevant to n16NN’s reported difficulty aggregating [Delak et al., 2007; Metzler et al., 2010].

The PRIME20-like model yielded far less evidence which can be favourably compared to existing hypotheses. The chains in these systems tended not to fold into each other, but rather sit next to each other. It may be that a greater number of n16N chains are required in a system for the most stable state to be one of chains folding together. However, the chains continued to show a far greater level of collapse than the PLUM* model or atomistic data [Brown et al., 2014], which may hinder aggregation.

There was little discrimination between interpeptide contacts, except for a slight preference in n16N-2 against interpeptide contacts forming with residues at the end of SD2 and start of SD3, which could be interpreted as loosely supporting the three-domain hypothesis. However, SD3 in the n16N-2 system was the least flexible subdomain.

5.4 The n16N-3 system

5.4.1 PLUM*

A REMD simulation of n16N-3 was carried out for 6.4 microseconds, using 55 replicas. The thermostatted temperatures were $T_i \in \{275.0, 275.9, 276.81, 277.71, 278.61, 279.52, 280.42, 281.32, 282.22, 283.13, 284.03, 284.93, 285.84, 286.74, 287.64, 288.55, 289.45, 290.35, 291.25, 292.16, 293.06, 293.96, 294.87, 295.77, 296.67, 297.58, 298.48, 299.38, 300.28, 301.19, 302.09, 302.99, 303.9, 304.8, 305.7, 306.61, 307.51, 308.41, 309.31, 310.22, 311.17, 312.21, 313.35, 314.64, 316.15, 318.09, 321.15, 324.77, 328.38, 331.99, 335.6, 339.21, 342.82, 346.43, 350.0\}$ K.

42799 frames were used for analysis of the system. The geometric clustering analysis was carried out with an RMSD cut-off of 0.8 nm, and the results show a tendency towards a single, strongly favoured top cluster has emerged. The top four clusters had populations of 7.0%, 1.5%, 1.3% and 0.93%, and these are shown in fig. 5.9. The dominant core type seen in n16N-2 PLUM* (fig. 5.1a), involving β -hairpins starting at SD1, has no analogue in the present system; instead, chains hook into each other, ending the hook with a turn at the end of SD2. In the top cluster, two chains hook the other chain together.

The regional clustering analysis showed very little change compared to the n16N-2 system. SD2 became even less flexible, its top cluster now representing a single strand of β -structure, turning at the SD2/SD3 boundary, as seen in all the top full-system clusters, with a population of 72.1%. The two other subdomains remained as before, SD3 staying the most flexible.

The degree to which each residue of the chain is involved in interpeptide interactions shows the same trend in the tripeptide system compared to the dipeptide; this is shown in fig. 5.11a. The increased level of hydrogen bonding in many residues in SD1 and SD2 highlights the lower level seen in residues K4 and K5 of SD1, and P15 and Y16 at the end of SD2, both of which are turning points in the chain in every full-system and regional top cluster.

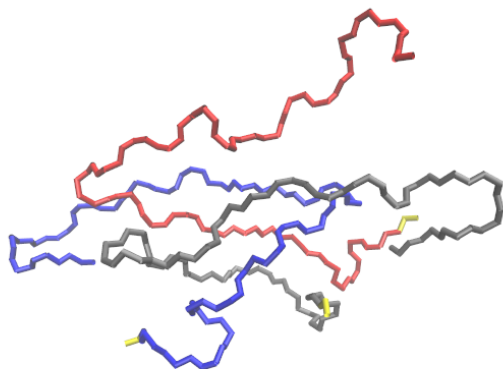
The Ramachandran plot of the system is shown in fig. 5.10, as well as a difference plot comparing to n16N-2 in PLUM*. The shift towards β -structure and away from α -structure has continued as the increased number of peptides has made β -strands running alongside each other increasingly favourable. The chain-hooking form of aggregation which dominates in this system is made of β -strands, and extended conformations which occupy β territory on a Ramachandran plot.

5.4.2 PRIME20-like

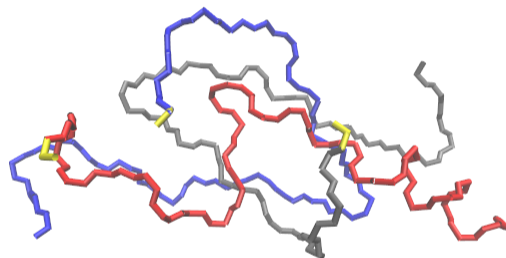
REMD simulations were carried out on the n16N-3 system in PRIME20-like in a large box. 48 replicas were used to evenly span the temperature range [0.105, 0.250], and each replica executed approximately 7×10^{10} events.

12338 trajectory snapshots at $T^* = 0.135$ were used for analysis. The full-chain clustering analysis in fig. 5.12 used an RMSD cut-off of 0.65 nm. The analysis shows that the chains have some ability to extend and fold together, rather than being purely fixed in a collapsed coil state. Unlike the PLUM* model, no strongly favoured structure has emerged yet. Additionally, no interpeptide interaction specificity on the level of subdomains or residues can be seen.

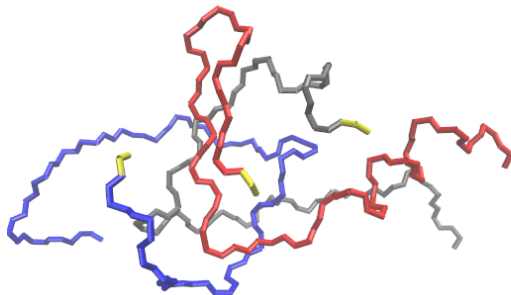
Fig. 5.11 shows the proportion of frames in which each residue is involved



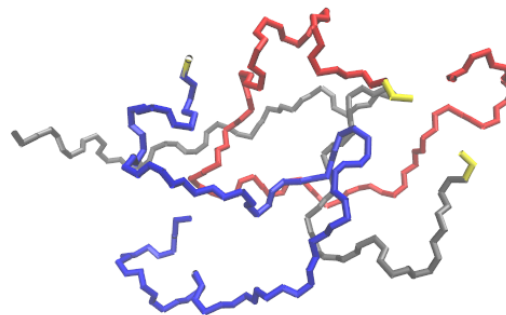
(a) **1 (7.0%)** The grey and red chains run with parallel β -strands from their K4 residues, turning at the end of SD2 and diverging in SD3. The blue chain encloses the two SD2 regions, again turning at the end of its SD2. Each chain's interpeptide interactions last approximately until its final tyrosine.



(b) **2 (1.5%)** The grey and blue chains form a core similar to n16N-2's fig. 5.1b. The third chain floats above, involving its SD2 in the core interpeptide interactions.



(c) **3 (1.3%)** The grey and blue chains loop around each other, turning at the end of SD2, and being involved in interpeptide interactions throughout SD2 and early SD3. The red chain is largely self-interacting, having a β -hairpin turning about residue G7, while SD3 features an α -helix. The red chain's SD2 is involved in interpeptide side-chain interactions with both other chains.



(d) **4 (0.93%)** Similar to 5.9c and 5.9b in having two chains loop around each other and a third nearby but outside of the loop. Turns occur consistently at residues K4 and K5, and at the end of SD2.

Figure 5.9: *The four top-occurring structures for the n16N-3 system in PLUM* at 300.0K. Each structure is labelled by its rank, with the percentage population given in brackets. N-termini are highlighted in yellow. At an RMSD cut-off of 0.8 nm, this analysis coarsely groups frames in which chains are similarly positioned, with little discrimination based on local intrapeptide secondary motifs.*

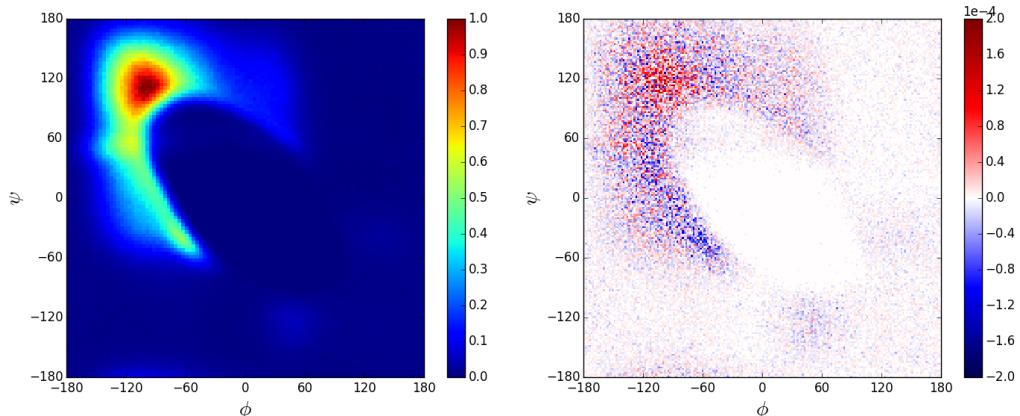


Figure 5.10: *Ramachandran plots of the n16N-3 system. **Left** shows a standard Ramachandran heat map. The system is strongly β -structure dominated. **Right** shows a difference heat map, comparing the (ϕ, ψ) coordinates visited with those of the n16N-2 system in PLUM*. The map is similar to the difference map of n16N-1 and n16N-2, shown in fig. 5.3, though the magnitude of the changes is far smaller.*

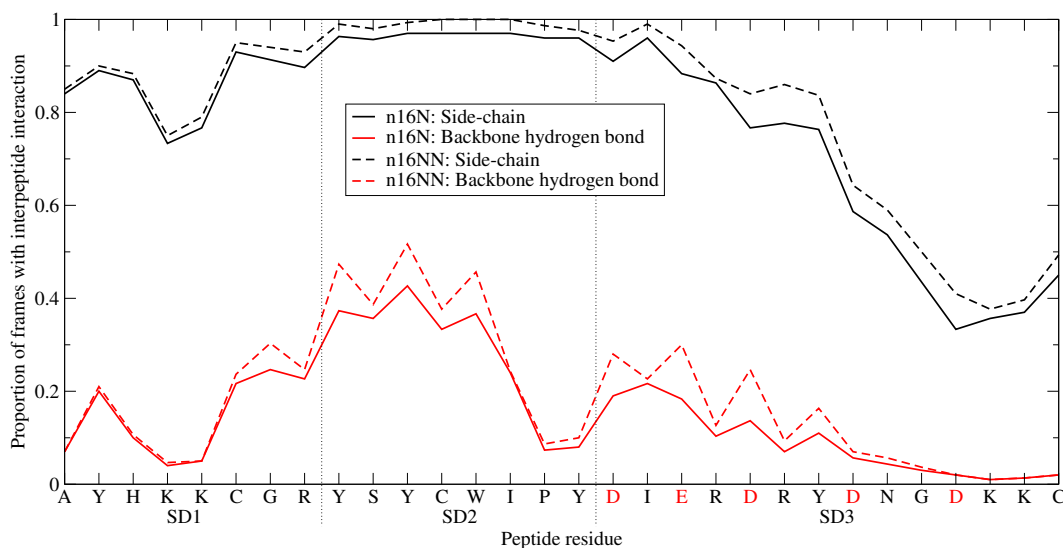
in an interpeptide interaction. The shift from a dimer system to a trimer system has not elicited any greater differentiation of regions from the chains on this metric. Although both lines' global minima exist around the SD2/SD3 boundary, the valley here is less pronounced than before. There is no evidence of distinct subdomains from this figure.

The Ramachandran plot of the system is shown in fig. 5.13, as well as a difference plot comparing to n16N-2 in PRIME20-like. The differences between this and smaller PRIME20-like systems remain minor compared to the large rearrangement seen with the PLUM* model.

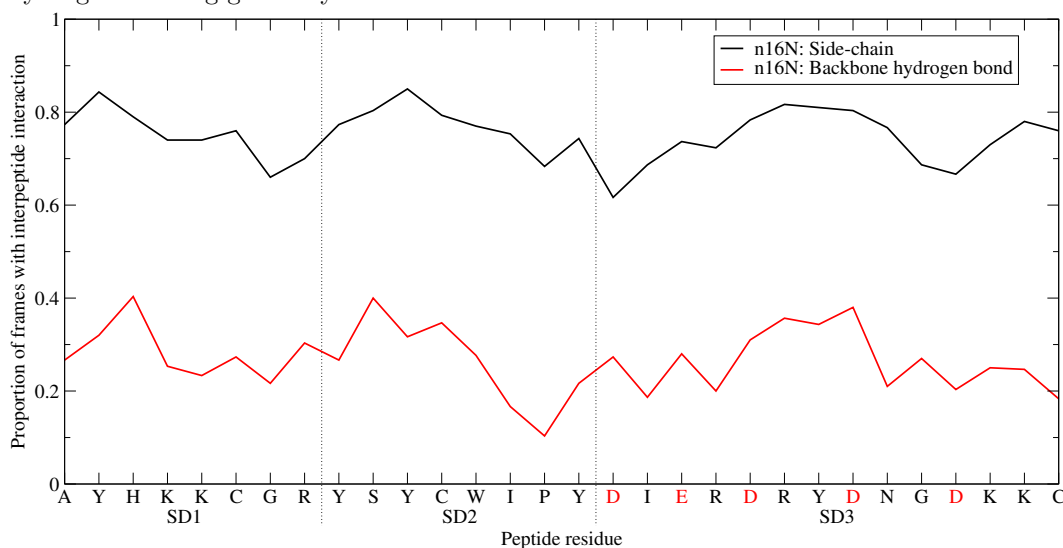
5.5 The n16NN-3 system

The n16NN-3 system was simulated in the PLUM* model only. The simulation time was 6.4 microseconds, and the simulation set-up of n16N-3 in PLUM* was repeated.

A full-system clustering analysis was carried out using 30734 frames. The top structure matches that of the n16N-3 system; two chains hooking the other one, with the geometry of fig. 5.9a. The stability of this geometric arrangement has risen; its frame population has moved from 7.0% to 12.3%. Fig. 5.11a shows that the change from n16N-3 to n16NN-3 does very nearly as much to stabilise interpeptide interaction outside of SD3 as inside it.

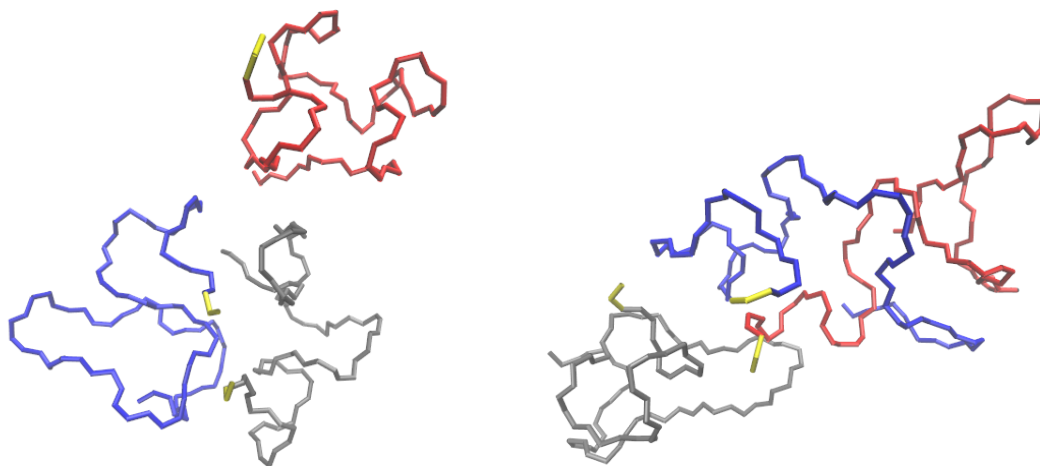


(a) **PLUM***: A clear distinction exists between the first two aggregation-aiding domains and the third; the tail. Compared to the dimer systems, fig. 5.2a, the side-chain interpeptide involvement is displaced upwards, however, the hydrogen bond line is increased by a factor instead, so the lower values at the middle of SD1 and end of SD2 have become more pronounced. In the clustering analysis, these are frequently seen to sacrifice the correct hydrogen bonding geometry to facilitate a turn.



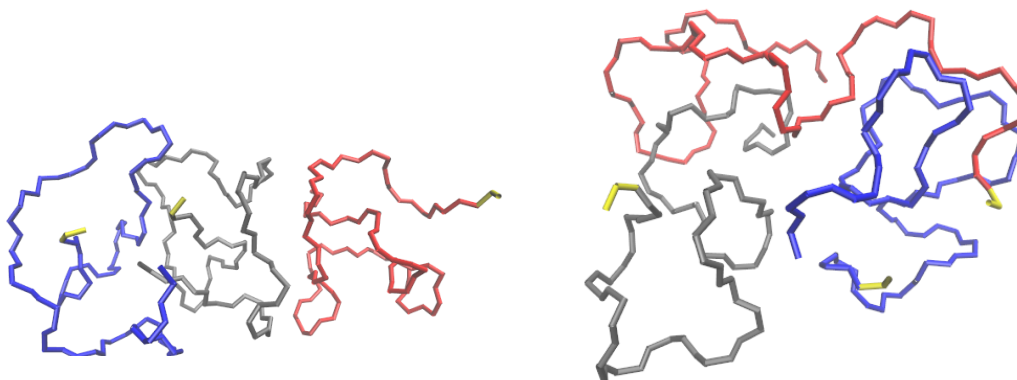
(b) **PRIME20-like**: No domain appears to be favoured for either hydrogen bond or side-chain based aggregation.

Figure 5.11: *The proportion of trajectory snapshots for which any given residue along the chain is involved in an interaction binding it to an other chain, in the $n16N-3$ and $n16NN-3$ systems. The $n16N$ residue sequence is shown on the x-axis; the red residues are replaced in $n16NN$ according to $D \rightarrow N$ and $E \rightarrow Q$. Interactions are divided into side-chain and hydrogen bond types; an interaction for glycine is not always applicable, and in these cases no data-point is plotted. Note that the disparate forms of interaction in each model make a comparison of the average of each line meaningless.*



(a) **1 (4.6%)** A large interface exists between the blue and grey chains, involving both ends of both chains. Where the three chains meet, SD1 and SD3 regions are heavily involved. No SD2 regions are involved in any interface.

(b) **2 (4.6%)** SD1 of the red chain is encompassed by the blue chain, and interacts primarily with the blue's SD1 and SD2. SD1 and SD2 of the grey chain are at the interface, interacting primarily with each other chain's SD1.



(c) **3 (4.4%)** The grey chain takes on a flat, wide shape to provide a large interface with both other chains. There is no clear subdomain or residue specificity in the way the chains aggregate.

(d) **4 (4.2%)** The red chain is spread out, and its N-terminal end interacts with the blue chain, while its C-terminal interacts with the C-terminal end of the grey chain.

Figure 5.12: *The four top-occurring structures for the n16N-3 system in the PRIME20-like model at $T^* = 0.135$. Each structure is labelled by its rank, with the percentage population given in brackets. N-termini are highlighted in yellow.*

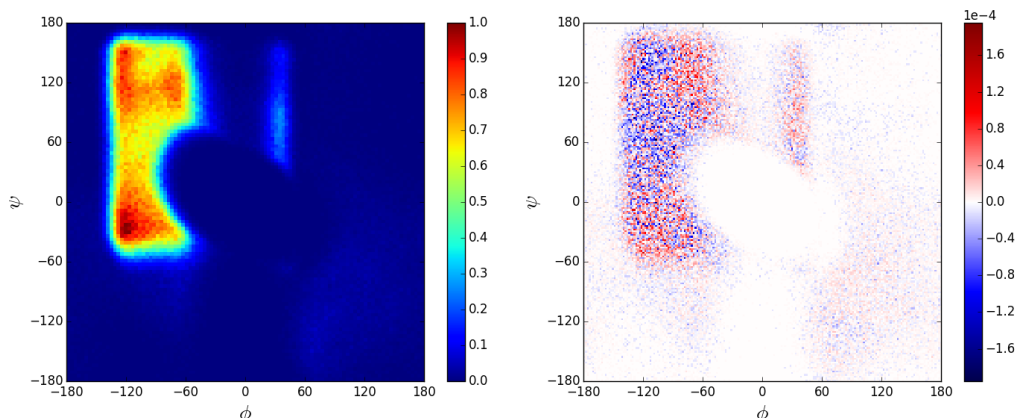


Figure 5.13: *Ramachandran plots of the n16N-3 system in PRIME20-like. **Left** shows a standard Ramachandran heat map. The global maximum is now the turn structure in the bottom left of the accessible region. Two significant, broad peaks also exist at low and high values of ϕ in the β -structure domain. **Right** shows a difference heat map, comparing the (ϕ, ψ) coordinates visited with those of the n16N-2 system in PRIME20-like. Within the top left quadrant, the peak for lower values of ϕ has weakened, while the peak for higher values has increased, as it did between n16N-1 and n16N-2. The trends concerning the bottom left quadrant have reversed, however.*

5.6 Discussion of trimer systems

The PRIME20-like modelling of n16N-3 was startlingly homogeneous in behaviour by residue or by subdomain along the chain. Despite a sophisticated parametrisation algorithm for side-chain to side-chain interactions (which did not have to be re-parameterised in the current work), with free parameters in energy and length, it seems unlikely that the simulation results would show an interesting difference if the sequence were scrambled.

The background evidence from atomistic simulation makes it unlikely that the PRIME20-like model is portraying n16N accurately. The delta in radius of gyration, first seen in the monomer system, is a huge anomaly which shows that the PRIME20-like simulations are deviating from predictions. It may be that PRIME20-like needs its interaction strengths tuning down in order for disordered systems to be able to properly expand, similar to PLUM*. However, the fact that the radius of gyration hardly grows in response to large temperature increases (see section 4.2.2) suggests that the problem lies with the binary nature of the backbone potential.

The PLUM* simulations showed significant differences from the dimer simulations. The top dimer n16N structure, with SD1 and SD2 forming β -hairpins, was

not favourable enough in the trimer system to reappear. Instead, chains began to interlink, as seen in all of the top structures. The propensity for β -structure grew at the expense of α -helix structure, though the same change was greater in magnitude between the monomer and dimer systems.

However, several features of the n16N chain have been preserved: A region similar to SD2, but perhaps impinging on SD1 and ending at the proline residue, is primarily responsible for aggregation. The first five residues do not appear essential for this. The C-terminal half of the chain is not directly involved in aggregation, and is the most highly disordered and flexible region. Folding is facilitated by turns at the ends of this aggregation domain, on residues K4 and K5 in SD1 and P15 and Y16 in SD2. The change from n16N to n16NN increases stability of the aggregates without changing the geometry of the top structures. The SD3 region, which is mutated in n16NN, may have a role in ensuring marginal stability of the aggregates, which may be a prerequisite of the macromolecular assembly coming together.

5.7 The n16N-6 system

The n16N-6 system is the largest simulation carried out to study the n16N peptide, giving it the best vantage point for clues about how large-scale n16N systems would assemble.

5.7.1 PLUM*

A REMD simulation of n16N-6 was carried out, lasting for 2.5 microseconds in each of 120 replicas. 94 replica temperatures were spaced out at even intervals of 0.59 K from 275 K to 330 K, after which the spacing was linearly increased as a function of temperature, reaching double its initial spacing at the maximum temperature of 350 K.

4157 frames were used for geometric clustering with an RMSD cut-off of 1.13 nm. The top four clusters have populations of 3.1%, 2.7%, 2.3% and 2.1%, and the top two of them are shown in fig. 5.14. The greatest number of strands in a β -sheet that can be stable appears to be four. Besides the top two, other popular clusters involve three to four chains in a β -sheet, with varying levels of β -structure and disorder among the extra chains. The extra chains are always found on the same side of the β -sheet, opposite to the N-termini. In the four-chain-sheet case, extra chains do not seem to conform to the previously seen dimer conformations of fig. 5.1, but instead run perpendicular to the primary β -sheet, towards the turn at

the residue P15. These observations suggest that n16N-8 may favour a layout of two four-chain β -sheets in a cross-sheet formation.

Fig. 5.15 is a Ramachandran heat map and difference heat map. These show that increasing the number of peptides in the system continues to increase order, decreasing the spread of dihedral angle pairs and increasing the magnitude of the β -peak and other minor peaks which facilitate n16N-6's favoured structures.

A regional clustering analysis was carried out as usual with an RMSD cut-off of 0.2 nm using same-length representations of proposed subdomains SD1, SD2 and SD3. SD2's top cluster is a β -strand as in n16N-3, but its population fell from 72.1% in n16N-3 to 63.7%. This may be unexpected, as the Ramachandran β -peak became stronger in n16N-6. However, n16N-6 typically has two chains not involved in the main β -sheet, which can be structured as forms of β -strands that are not geometrically similar enough to the main β -sheet structure to be clustered together. SD2's top three clusters are all β -strands with differing turns at the ends. SD1 favoured the looping structure seen in the top clusters of fig. 5.14 with a population of 55.6%. SD3's top structures had populations of 48.1%, 20.6% and 12.7%. The third cluster is a β -hairpin structure, while the other two are extended conformations which include both disorder and β -structure. This still provides a comparison point for the level of flexibility in the subdomain, however.

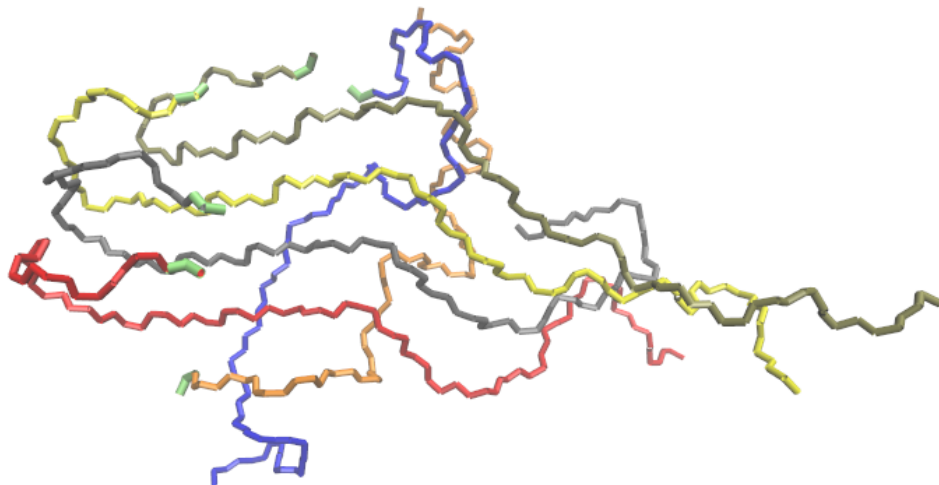
The proportion of interpeptide interactions has again been analysed, and this data is presented in fig. 5.16a. The graph supports the hypothesis of an SD2-like aggregation domain and an SD3 'free tail' domain, which has consistently been shown to have merit in n16N systems examined in PLUM*.

5.7.2 PRIME20-like

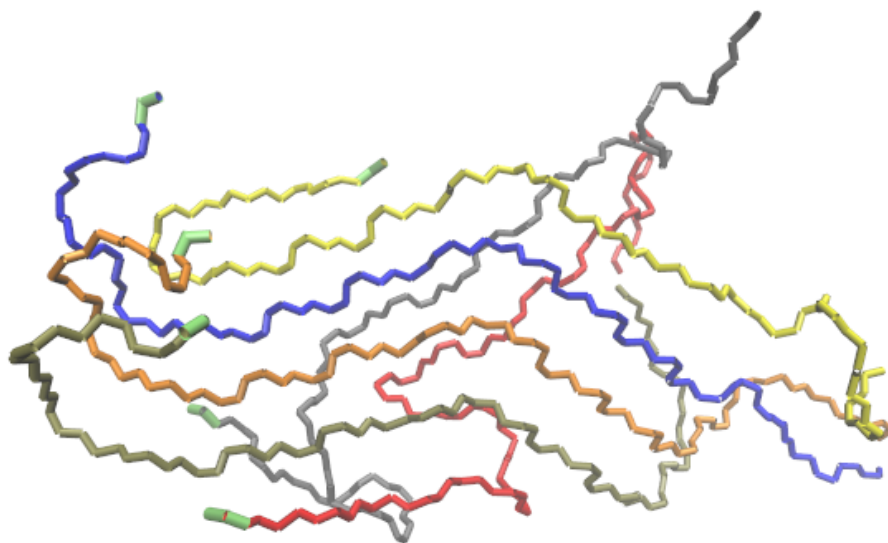
REMD simulations on the n16N-6 model were carried out with 48 replicas, evenly spaced in the reduced temperature range [0.135, 0.250]. Each replica executed approximately 1.4×10^{10} events.

2180 frame snapshots were used at $T^* = 0.135$ for geometric clustering analysis. At an RMSD cut-off of 0.8 nm, the top four clusters had populations of 4.7%, 4.0%, 3.6% and 3.1%. The top two are presented in fig. 5.17. The structures remain extremely disordered and collapsed. Nonetheless, they differ from the dimer and trimer systems by being strongly involved with each other. The Ramachandran plots of the trajectory, shown in fig. 5.18, reinforce the observation of great disorder in this system, and show that little has changed on the level of secondary structure.

The analysis of propensity for interpeptide interactions by residue, fig. 5.16b, shows the system has even increased in homogeneity, on general propensity for



(a) **1 (3.1%)** Four strands form a parallel β -sheet. Except for the last chain (golden, behind yellow), their SD1 regions are not in β -hairpins, but instead each hovers over the start of the *next* chain's SD2. Each strand is disrupted at the proline residue, and the turn occurs at approximately the SD2/SD3 boundary. After this, some hydrogen bonding exists but the chains incrementally become more free. Below the β -sheet, two chains aligned perpendicular to it and anti-parallel to each other are highly disordered.



(b) **2 (2.7%)** Similar arrangement to fig. 5.14a with more order. After the first turn at the proline residue, β -sheet structure continues almost undisrupted until the final tyrosine; Y23. The red chain's SD1 is involved in the β -sheet, while its SD3 up to residue Y23 forms a β -strand with the same region of the grey chain.

Figure 5.14: *The two top-occurring structures for the n16N-6 system in PLUM* at 300.0 K. Each structure is labelled by its rank, with the percentage population given in brackets. N-termini are highlighted in lime.*

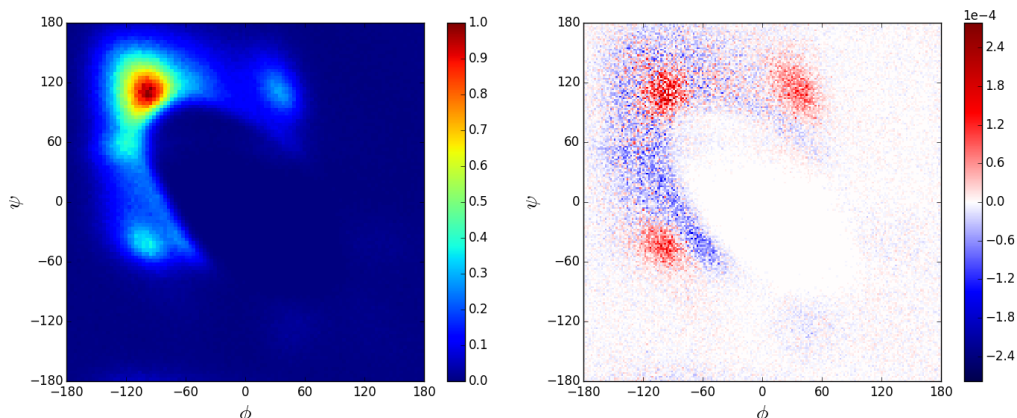
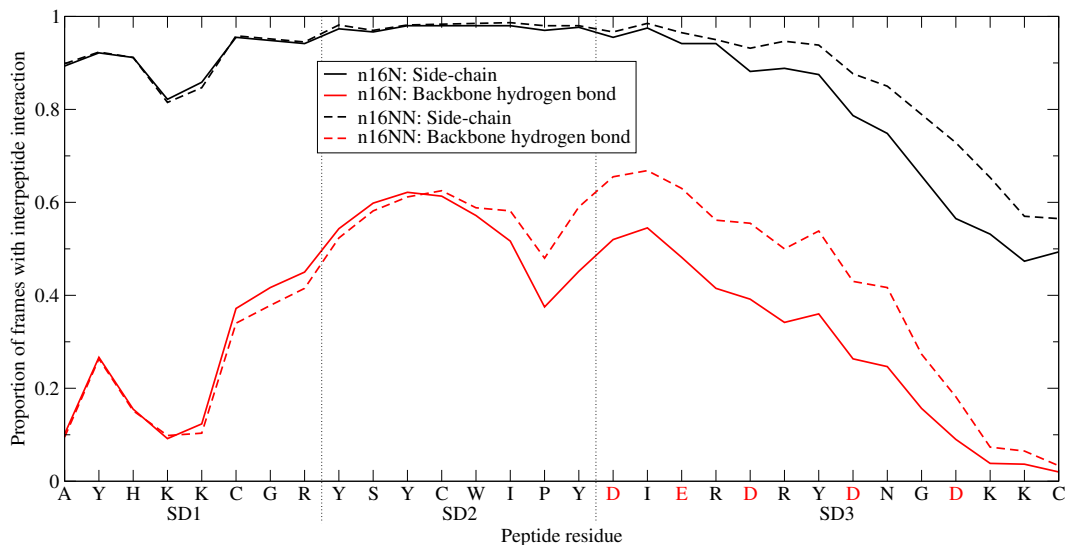


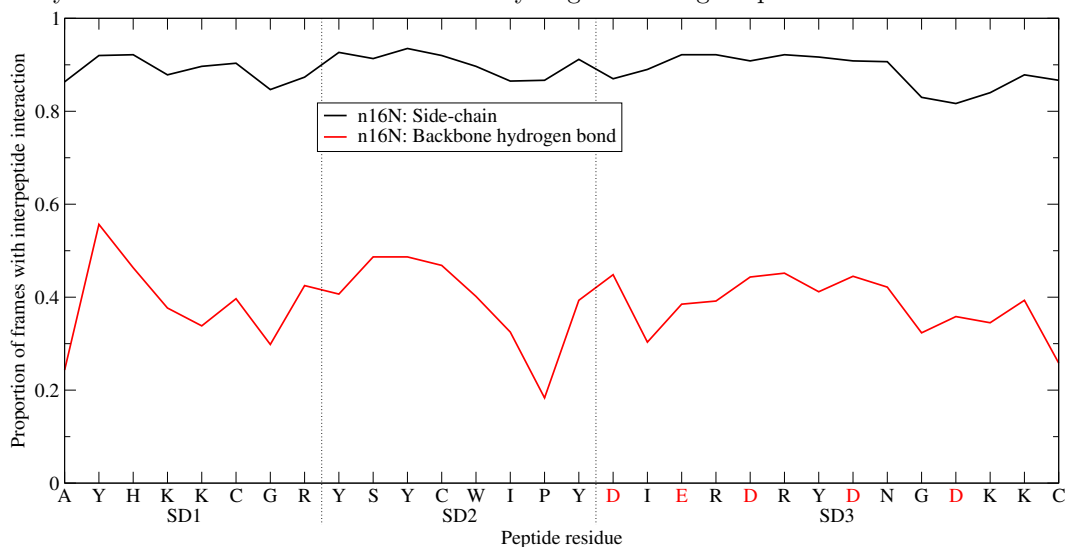
Figure 5.15: *Ramachandran plots of the n16N-6 system. **Left** shows a standard Ramachandran heat map. The plot is dominated by the β -structure peak, which is twice the magnitude of any other peak. A wide range of other angle coordinates are accessible. **Right** shows a difference heat map, revealing the difference between the normalised n16N-6 and n16N-3 systems. Continuing the trend from n16N-1 to n16N-2 to n16N-3, most accessible regions and particularly the α -helix region have drained, while the β -structure peak has risen. However, new peaks can be seen growing at coordinates $(-50^\circ, -100^\circ)$ and $(40^\circ, 110^\circ)$. These dihedral angle pairs occur in SD1 and SD2 at the start and end of β -strands.*

interchain interactions. However, the n16N-6 system's proclivity for interpeptide interaction has made a more detailed form of analysis worth performing. In fig. 5.19, a 2D heat map of residue-residue interpeptide interactions is given. This shows *which residues* of other chains any given residue is likely to interact with. It reveals that the model does indeed have an ability to distinguish between regions of the chain; all regions of the chain preferentially interact with the N-terminal region of other chains, and this preference is strongest for the N-terminal region itself. The global peak is for residue Y2 interacting with itself, but the largest island of interaction is in the hypothesised aggregation domain named subdomain SD2.

The regional clustering analysis, which uses same-length representations of hypothesised subdomains SD1, SD2 and SD3, returned similar results to the dimer system, in which SD3 is the least flexible subdomain. SD1's top cluster populations were **4.2%**, **2.2%** and **2.0%**. SD2's were **2.8%**, **2.1%** and **1.9%**. SD3's were **5.3%**, **3.5%** and **3.4%**. SD1 and SD3's top structures were β -hairpins, turning about the K4, K5 and N25, G26 subsequences, respectively.

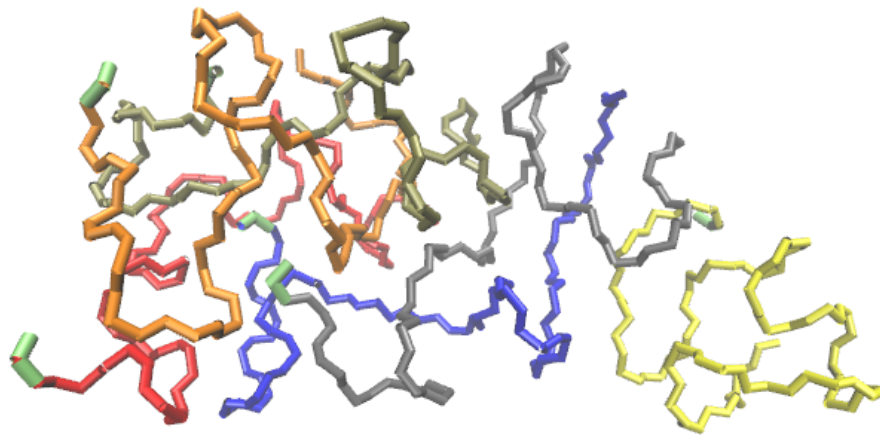


(a) **PLUM***: The hydrogen bond line shows that the most important region for peptide aggregation is similar to SD2, displaced one to two residues left (i.e. towards SD1). Arguably, the start of SD3 may also be included. The trend is very similar to n16N-3, despite different top geometric clusters. SD3's length is just sufficient for its final residues to be far enough away from the bulk of the cluster that its hydrogen bonding drops almost to 0.

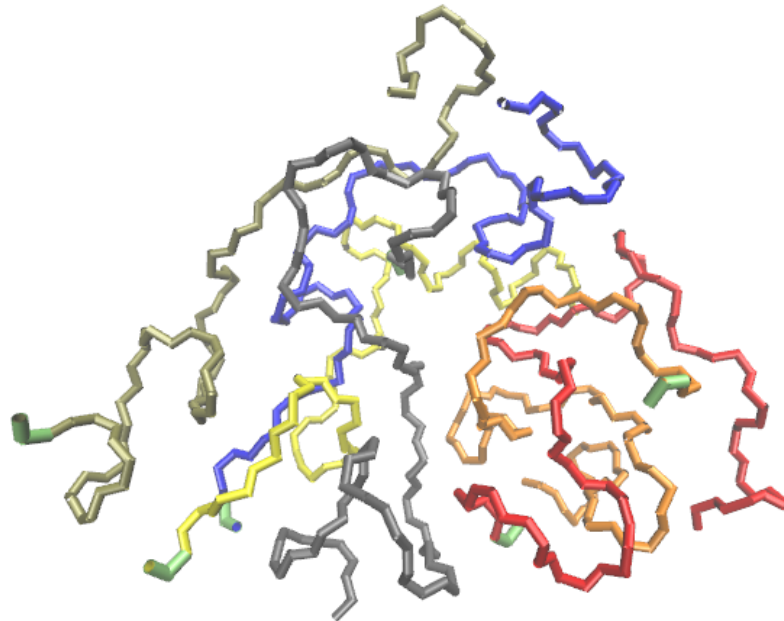


(b) **PRIME20-like**: The side-chain line is extremely flat, revealing essentially no preference in how the chains aggregate. The hydrogen bonding line shows a little variation. The SD2 region has a similar form to that of 5.16a, though that is largely a predetermined result of proline's nitrogen having its hydrogen bonding capability disabled. No n16NN-6 simulation was carried out in this model.

Figure 5.16: *The proportion of trajectory snapshots for which any given residue along the chain is involved in an interaction binding it to another chain, in the n16N-6 and n16NN-6 systems. The n16N residue sequence is shown on the x-axis; the red residues are replaced in n16NN according to $D \rightarrow N$ and $E \rightarrow Q$. Interactions are divided into side-chain and hydrogen bond types; an interaction for glycine is not always applicable, and in these cases no data-point is plotted. Note that the disparate forms of interaction in each model make a comparison of the average of each line meaningless.*



(a) **1 (4.7%)** The yellow chain sits on its own and primarily self-interacts. The other five chains are highly involved with each other, though not through clearly recognisable motifs like β -strands. N-terminal chain ends are found close together in this structure. There is no obvious order to the structure.



(b) **2 (4.0%)** All chains are interwoven; none are primarily self-involved. No clear common secondary structural motifs are visible, besides the yellow chain's C-terminal α -helix. There is very little order to the coil.

Figure 5.17: *The two top-occurring structures for the n16N-6 system in PRIME20-like at $T^* = 0.135$. Each structure is labelled by its rank, with the percentage population given in brackets. N-termini are highlighted in lime.*

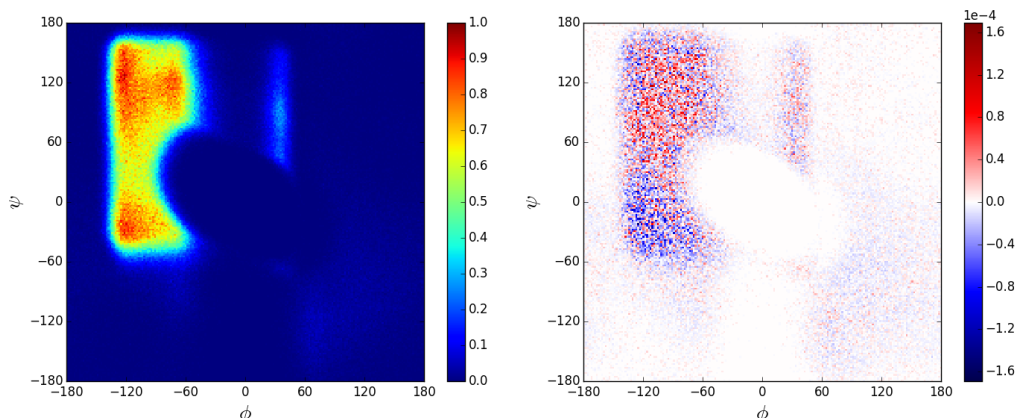


Figure 5.18: *Ramachandran plots of the n16N-6 system in PRIME20-like. Left shows a standard Ramachandran heat map. Within the island of sterically accessible dihedral angles, no coordinates are strongly favoured or disfavoured. Right shows a difference heat map, comparing the (ϕ, ψ) coordinates visited with those of the n16N-3 system in PRIME20-like. While the changes seen here are reversed compared to the changes between the dimer and trimer n16N systems, the magnitude of changes observed remains very low.*

5.8 The n16NN-6 system

The n16NN-6 system was simulated in the PLUM* model only. The simulation time was 3.3 microseconds, and the simulation set-up of n16N-6 in PLUM* was repeated.

The interpeptide interaction analysis on a per-residue basis of fig. 5.16a deviates from previous results. In the dimer and trimer systems, the change from n16N to n16NN caused *full-system* interpeptide interaction proportions to rise, that is, SD1 and SD2 were affected similarly to SD3. However, the present system has C-terminal residues' proportions rising while N-terminal residues stay at the same level or decline slightly.

5543 frame snapshots were saved for use in a clustering analysis. An RMSD cut-off of 1.13 nm was used, making the results comparable to n16N-6 in PLUM*. The top clusters showed far more stability than n16N-6 in PLUM*, having populations of 10.1%, 6.0%, 5.0% and 3.8%. As in n16N-6, the top clusters all have a three or four-chain β -sheet with extra chains found near the β -sheet's SD2 region, on the side of the sheet opposite the N-termini. However, the top cluster now features the two extra chains in a more tightly packed conformation, illustrated in fig. 5.20.

The subdomain clustering analysis revealed that neither SD1 or SD2 were significantly changed compared to n16N, their top three geometric clusters all falling within 2.0% of their values in n16N-6. Only SD3 showed significant change, the

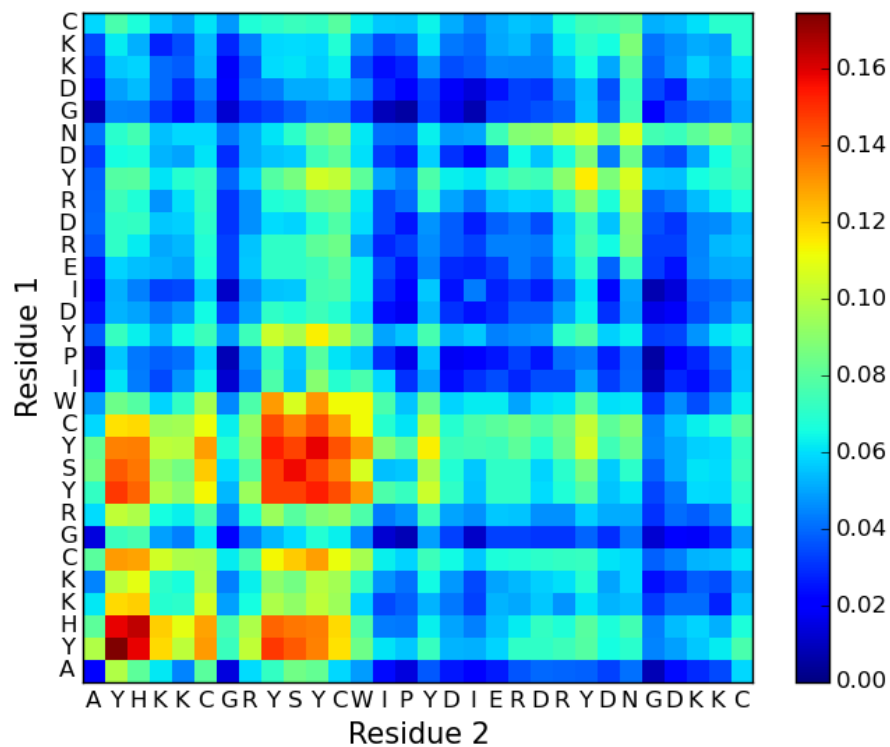


Figure 5.19: A heat map illustrating the frequency of interpeptide residue-residue contacts in the PRIME20-like model's $n16N-6$ system. The dataset plotted includes all non-bonded interactions except for backbone hydrogen bonding. A value of 1.0 would indicate two residues interacting in as many pairs as possible (15 pairs for a 6-chain system) in every frame. On this scale, a value of 0.16 implies an average of 2.4 corresponding residue pairs are in contact.

stability of the extended or β -strand conformations of its top two clusters growing by approximately 4.0% each, while the third-place β -hairpin dropped from 12.7% to 5.6%.

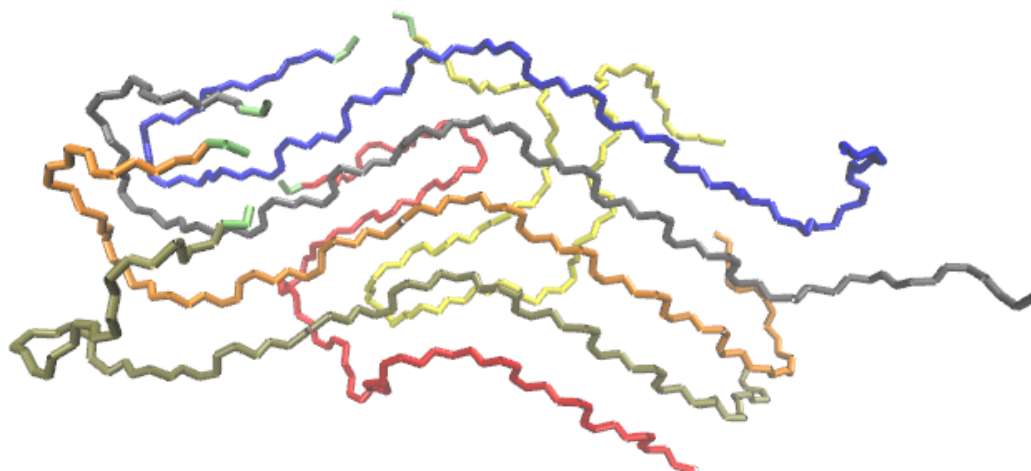


Figure 5.20: *The topmost populated geometric cluster of the n16NN-6 system in PLUM*, with a population of 10.1%. Four chains form a parallel β -sheet, while the other two are tightly packed in a hybrid β -strand/ β -hairpin conformation which involves both inter- and intra-peptide hydrogen bonding. SD3 of the red chain joins the β -sheet, which would be very unusual for n16N.*

5.9 Discussion of hexamer systems

In PLUM*, the n16N-6 system took on a new favoured structure, based around the SD2s of three or four chains coming together to create a β -sheet. The trend of β -structures growing in dominance as system size increases has persisted. The stable β -sheet top structures which dominated the n16N-6 simulation are the first top structures seen for any system size which could conceivably be scaled and could manifest in bulk without total rearrangement.

As in all n16N PLUM* simulations, the pattern of SD2 as the aggregation domain and SD3 as the flexible and disordered tail remained. Nothing in the n16N-6 simulations suggests particular function of SD1; it does not seem to be crucially involved in aggregation, nor does it possess the properties of SD3, charges and flexibility, which would suggest a role with ions. The staple of top structures having

turns around the residues K4 and K5, and P15 and Y16, recurs in the n16N-6 and n16NN-6 systems.

The mutant peptide n16NN, which has been observed to aggregate abnormally or not at all [Delak et al., 2007; Metzler et al., 2010], features top geometric clusters which are the same as n16N in the N-terminal, aggregation-enabling half. Unlike the monomer, dimer and trimer systems, the SD1 and SD2 regions which make up the N-terminal half of the chain are not more stable in n16NN than n16N, but instead about equally so. This challenges the hypothesis which emerged from the dimer and trimer systems that n16NN decreases full-chain flexibility, which may be necessary for macromolecular assembly. The n16NN-6 top structure in fig. 5.20 shows SD3 becoming involved with the other subdomains in interpeptide interactions, and this is confirmed by the comparison of interpeptide interactions by residue (fig. 5.16a). It is also seen in the smaller systems. It may simply be that the SD3 domain of n16NN stops aggregation by outcompeting other chains for interpeptide interactions with SD2.

The PRIME20-like simulation of the n16N-6 system had similar results to smaller systems, in that the random coils which resulted were entirely collapsed and disordered. The graph presented on interpeptide interactions per residue showed that the random coil remained uniformly bound to itself, no portion of the chain becoming a free tail or interacting less than other portions. A subdomain clustering analysis showed that SD3 was still the least flexible subdomain. These aspects of the results were largely in violation of the predicted clustering behaviour of the system, and the collapsed nature of the structures is particularly troubling.

The present system also differed from smaller systems, in that chains ceased to prefer self-interaction, and folded together to a significant degree. An analysis of which residues on other chains each residue prefers to interact with proved extremely important, because it revealed that there are huge preferences and there is a kind of order to n16N-6's structures. SD2-SD2 interactions are far more common than any other subdomain-to-subdomain interaction, which does support the subdomain hypothesis. The fact that the PRIME20-like model perhaps gets this aspect of the system correct, while secondary structure and the collapsed nature of the coil appear completely wrong, suggests that the side-chain and non-bonded aspects of the model are stronger than the backbone steric and/or hydrogen bonding aspects.

5.10 Summary

The PLUM* coarse-grained model was used to simulate systems of 2, 3 and 6 units of the peptides n16N and n16NN. REMD simulations were used, with replica counts of 30, 55 and 120 spanning the temperature range [275, 350] K.

The original PLUM model was also used to simulate n16N-2, and the result confirmed the need for the change in hydrogen bond strength which led to the PLUM* model. n16N-2 in PLUM yielded two chains both with α -helix structure and with a kink in the middle, exactly as in the top structure of n16N-1, which sat next to each other and interacted almost exclusively via their side-chains. After the small tweak to the PLUM model, chain multiplicity brought a remarkable degree of region-specific differentiation and function out of n16N.

The subdomain hypothesis, which categorises regions of the n16N chain by local function and is delineated in table 1.3, was repeatedly validated with respect to the aggregation behaviour of SD2 and the flexibility of SD3. The work elucidated no clear role for SD1, but it is possible that the simulation would require other elements, such as a suitable surface, for SD1 to activate.

As system size increased, so did propensity for β -structure, and the largest simulations produced four-chain β -sheets. β -sheets are scalable with more chains, so there is the possibility that this would stay the means of aggregation in larger systems.

n16NN simulations were carried out to check for differences in the properties of the two systems. An inability of n16NN to aggregate has been experimentally observed, and it was hoped that simulation might give a hint about why. In the dimer and trimer systems, n16NN had greater full-system stability than n16N and maintained its interpeptide interactions in a higher proportion of trajectory snapshots. However, this was no longer true in the hexamer systems. One consistent trend observed was a far higher probability in n16NN than n16N for the SD3 to become involved with the aggregating SD2 regions. Therefore, it was proposed that n16NN's SD3 may interfere with aggregation by blocking further interpeptide interactions.

The PRIME20-like model was used to simulate 2, 3 and 6 units of the peptide n16N. It was also used to simulate 2 units of n16NN. REMD simulations were used, with replica counts of 30, 48 and 60 spanning the temperature range [0.105, 0.250] linearly. The reduced temperature $T^* = 0.135$ was used for analysis, though examinations of datasets at other temperatures revealed that structure was very similar over a wide range.

Chains of n16N and n16NN were fully collapsed throughout, and remained

so even at high temperatures. Chains folded into tightly packed random coils which were barren of recognisable secondary motifs. In no system were strongly favoured top structures observed. In dimer and trimer systems, the collapsed coils were primarily self-interacting, typically sitting next to each other with mainly side-chain interpeptide interactions.

Repeatedly, Ramachandran plots showed peaks for β -region structure and for a turn motif with coordinates of (121, 28), both of which were also seen in the monomer system. Contrary to hypotheses asserting that n16N may take on some degree of order as it assembles in the presence of other chains, increasing chain number brought about greater flatness in the island of sterically allowed coordinates on the Ramachandran plots.

The hexamer system favoured interpeptide interaction far more, the chains intertwining in the four top structures. The chains remained fully disordered and maximally collapsed, but an analysis of which residues interact with which others revealed a strong preference for interactions to occur between SD2 and SD2, SD1 and SD1, and SD1 and SD2. This was a strong sign that the PRIME20-like model can differentiate between regions of the n16N peptide. This result agreed with the subdomain hypothesis of [Brown et al., 2014].

Chapter 6

Conclusions

This project set out to advance the role of simulation to study intrinsically disordered proteins, by producing an accelerated simulation methodology which works for them. The already ubiquitous replica exchange method was selected for its general availability and applicability as the accelerated sampling method of choice. Two coarse-grained models were selected to study the simulation targets n16N and n16NN. In the following sections, the models' success in simulating the peptides, and what has been learned about the peptides themselves, will be evaluated. The PRIME20-like and PLUM* simulations produced incommensurate trajectories with very little in common, therefore, they will be analysed separately.

6.1 PLUM* model

The PLUM* model is the coarse-grained PLUM model [Bereau and Deserno, 2009], with a decrease of 5.5% to the backbone hydrogen bonding strength. This was found in section 4.1.1 to cause n16N-1 simulations to better match existing atomistic data. The model was used to simulate systems of 1, 2, 3 and 6 units of the intrinsically disordered peptide n16N and its mutant n16NN. It was also unsuccessfully used to simulate the designed S1 peptide, which is known to form a polyproline II helix as its native state.

6.1.1 Simulation results

Each simulation of n16N systems at different chain numbers produced different top geometric clusters, without recurrence of tertiary structure between system sizes. This is to be expected at such small system sizes, but means that, very possibly, the tertiary structure evinced from n16N in these systems does not represent bulk

behaviour. However, one of the consistent trends was an increase in β -structure as number of chains rose. β -structure was the dominant aggregation form in multiple-chain systems. Systems of two and three units were rich in parallel β -strands, and these evolved to parallel β -sheets in the n16N-6 system. The spare two chains in the n16N-6 system, when four chains were in a β -sheet, favoured a conformation which suggested a capacity for layering of β -sheets together.

CD spectral evidence shows that n16N oligomers are composed of 46% β -structure and 54% random coil structure, and that these proportions are similar to the monomeric state [Amos et al., 2011]. The large PLUM* simulations showed a similar balance of structure to this, the N-terminal half of the chain likely to be in β -structure, and the C-terminal half likely to be disordered. Conversely, the monomeric state was significantly more α -structured, even after turning down the backbone hydrogen bond strength. PLUM* was tuned to match the α -helicity of a CHARMM22* simulation [Brown et al., 2014] by matching the population of areas of the Ramachandran plot, but comparing the most populated geometric clusters of n16N in PLUM* and in CHARMM22* reveals that this did not perfectly eliminate all excess stable α -helix manifestation in the monomer. However, in multiplicity, the adjustment lead to far more interesting results.

Every simulation of n16N peptides broadly agreed with the subdomain hypothesis [Brown et al., 2014] outlined in table 1.3. In every system including n16N-1, SD3 possessed the greatest conformational freedom. In multi-peptide simulations, SD2 was most involved in interpeptide interactions, while involvement in interpeptide interactions declined approximately linearly from the start of SD3 onwards. Top geometric clusters repeatedly had Y23 as the last residue with interpeptide hydrogen bonding. In the n16N-1 simulation, SD1 and SD2 were not distinguishable. However, a difference in involvement in aggregation appears in the n16N-2 system, and grows in the larger systems. The precise ‘interpeptide stabilisation’ domain is not clearly defined by results in the present project. Arguably, it begins before the SD2 domain starts, and, according to top structures from multiple systems, lasts as far as residue Y23. The precise length of the flexible tail left over is probably sensitively dependent on ambient conditions which the PLUM* model cannot capture.

SD1’s two K residues manifested as extremely structure-disruptive, usually placing a lower limit on the range of any interpeptide structural motif. The two Ks also facilitated turns, leading to rather weakly bound β -hairpins or side-chain mediated interactions with SD2.

Simulations of the mutant of n16N known as n16NN were carried out at the same system sizes. Since it is known that n16NN cannot assemble normally

[Delak et al., 2007; Metzler et al., 2010], it was hoped that these simulations would bring to light the important differences between the chains. Monomers of n16NN were more stable overall than n16N monomers, and in the dimer and trimer cases, full-chain stability for the top structures was greater in n16NN, as was the stability of interpeptide interactions along the full chain. In the hexamer systems, the N-terminal half of the chain was no more stable in n16NN than n16N. Consistently in all systems, the n16NN SD3 was far more prone to interact with the rest of the chain and with other chains than the n16N SD3. Based on this, it is possible that the n16NN C-terminal tail disrupts aggregation by competing for interactions and sterically blocking other chains.

The PLUM* results in this project have been extremely impressive, the different residues in the chain leading to the emergence of three clearly distinct areas of different function, based on a simple hydrophobic interaction scale with mixing rules. These results paint an optimistic picture of the future role of coarse-grained models in the difficult case of intrinsically disordered peptides and proteins. In both n16N studies, and in general, continual comparison of results with atomistic data will be necessary to refine models and validate results.

6.1.2 Further work

The attempted simulation of the S1 peptide was the single clear failure for the PLUM and PLUM* models. This revealed a deficiency in the model relating to proline. The PLUM* model was designed to have success finding the most common secondary structural motifs, which are stabilised by strong backbone hydrogen bonds. However, the polyproline II helix that S1 correctly folds into is not stabilised by backbone hydrogen bonds, but instead by side-chain interactions. It would be interesting to see whether improvements to the realism of the proline residue, especially limiting the ϕ dihedral angle to approximately -60° , would serve to enable correct S1 folding. A change in side-chain interactions may also be necessary. In n16N, the proline residue is structure-disruptive and often the site of turns, so improving the realism of proline's dihedral angular potentials may be particularly beneficial for the successful modelling of n16N systems.

Larger system sizes of n16N would be of interest and can be carried out with ease. The reported n16N-6 and n16NN-6 simulations used 120 cores on the local Warwick University cluster Minerva for just 384 hours and 576 hours respectively. Of the cores used, 43 were below 300 K, at temperatures which were not referred to in the final analysis. Optimisation of simulation strategy, and use of national computational facilities, would enable larger-scale simulations to be performed. These

simulations could answer questions about whether and how hypothesised multi-layer β -sheets come together, and would bring other system features such as ions and surfaces into scope.

To observe the interaction of PLUM* n16N systems with surfaces or ions, parametrisation would be required. Several existing atomistic forcefields already allow the combination of ions with proteins, so this feature could be brought into PLUM and PLUM* with multiple reference points for validation. Bringing a relevant surface, such as β -chitin, into the simulation would require a large project to create a suitable coarse-grained model.

Introducing ions would allow investigation of hypotheses regarding SD3's proposed "fly-casting" mechanism [Shoemaker et al., 2000; Brown et al., 2014]. The PLUM* model would also be tested for its ability to conform to the experimental results that n16N neither binds strongly to Ca^{2+} [Seto et al., 2014], nor has its conformational ensemble altered in these ions' presence [Collino and Evans, 2008]. Simulations orders of magnitude larger than the present system would be required to investigate the observation of n16N assemblies creating localised compartments of Ca^{2+} [Seto et al., 2014].

The more ambitious project of introducing a β -chitin model into PLUM* would make it possible to recreate a microcosm of the environment of n16N absorbed on β -chitin in the presence of Ca^{2+} , which has been observed to nucleate aragonite Keene et al. [2010a].

6.2 The PRIME20-like model

The PRIME20-like model is based on the PRIME20 model [Cheon et al., 2010]. Parameters which were not publicly available were filled in, and this is described in section 3.1. The model was used to simulate systems of 1, 2, 3 and 6 units of the intrinsically disordered peptide n16N and 1 and 2 units of its mutant n16NN.

6.2.1 Simulation results

The simulations of n16N defied predictions and produced unusual behaviour in multiple ways. All system sizes produced trajectories of extremely tightly packed chains, with no free tails or preferred aggregation domains. The systems featured almost no ordered secondary structural motifs whatsoever. All of the systems except n16N-6 predominantly formed intrapeptide interactions.

Diagnosing the problem requires more data and a larger variety of test systems, with the capability to isolate different aspects of the model. However, looking

at the reasons to be optimistic do provide a clue. The single-chain n16N system had promising results. The most common structures involved a buried SD2, maximising tyrosine interactions, as seen in atomistic simulations [Brown et al., 2014]. They also had an SD3 on the outside of the structure, though it was still bound to the coil. The six-chain system showed no greater tendency for SD2 domains to be in interpeptide interactions, but residues on the N-terminal half were found to be vastly more favoured as other residues' interaction partner than residues on the C-terminal half. The largest island of favoured interaction partners was in the centre of SD2, agreeing with the PLUM* model and the subdomain hypothesis. Therefore, the PRIME20-like model does strongly exhibit a form of residue specificity which may lead to reliable predictions when other errors in the model are fixed.

Secondary structure is chiefly derived from properties of the backbone, e.g. steric clashes, backbone-to-backbone hydrogen bonding and dihedral potentials. On the other hand, residue specificity *must* derive from side-chain interactions. Ignoring basic α -helix forming chains, PRIME20 has been used in published work to study chains of length 7 AA [Cheon et al., 2011], 10 AA [Wagoner et al., 2011], 6 AA [Wagoner et al., 2012] and 6 AA [Cheon et al., 2012]. It may be that the model is incapable of proper secondary structure with a chain of length 30 AA. On top of this, the steric properties of PRIME20 had to be overhauled to create the PRIME20-like model, and this occurred without significant testing.

The PRIME20-like model's simulations have been plagued by its inability to form secondary structure properly, and this has masked most of the model's finer properties, such as its carefully tuned side-chain interactions. The glimpses of the model's potential which have emerged from some datasets do spur hope for the model, but it will need improvements before re-evaluation for use in future studies.

6.2.2 Further work

Before PRIME20-like is used for more simulations of n16N or other IDPs, adjustment to the core parameter set is needed. It is not clear whether there is any simple fix for the backbone's structural behaviour, but a few suggestions can be made. Work should begin by investigating the steric freedom of the backbone; is it sufficient to allow all commonly observed secondary structure? If so, is the binary nature of the potential causing a lack of stiffness, and could a minimal set of multi-step potentials aid this? Testing against simple ordered and designed peptides would be extremely helpful.

The PRIME20-like model has been shown to behave successfully and very similarly to PRIME20 for short peptide fragments, as in section 3.1.2. The capability

to perform this kind of simulation with an open-source model may be useful to some researchers.

6.3 Outlook for coarse-grained models to study IDPs

The data presented in this project are sufficient to fully support the notion that coarse-grained models, at the resolution of four beads per residue, can augment the investigation of intrinsically disordered peptides. More specifically, the continuous-potential model PLUM [Bereau and Deserno, 2009] has been shown capable of demonstrating predicted behaviour of the peptide n16N in system sizes up to 6 chains. The model has the potential to simulate larger systems and to be enhanced to involve more features of the biomineralisation environment in which n16N exists. Pursuing this path could be instrumental to the development of our understanding of n16N and other intrinsically disordered proteins.

It remains to be seen whether the discontinuous-potential model PRIME20 [Cheon et al., 2010] and its derivatives can be made useful in the study of peptide chains exceeding ten residues. If it is possible, then this project has presented some indications that the model may be similarly useful to growing our understanding of intrinsically disordered proteins.

Bibliography

GNU General Public License, June 1991. URL <http://www.gnu.org/licenses/old-licenses/gpl-2.0.html>.

GNU General Public License, June 2007. URL <http://www.gnu.org/licenses/gpl.html>.

B. J. Alder and T. E. Wainwright. Studies in Molecular Dynamics. I. General Method. *The Journal of Chemical Physics*, 31(2):459–466, 1959. doi: 10.1063/1.1730376.

M. P. Allen and D. Quigley. Some comments on monte carlo and molecular dynamics methods. *Molecular Physics*, 111(22-23):3442–3447, 2013. doi: 10.1080/00268976.2013.817623.

M. P. Allen and D. J. Tildesley. *Computer simulation of liquids*. Oxford University Press, 1987.

F. F. Amos, C. B. Ponce, and J. S. Evans. Formation of framework nacre polypeptide supramolecular assemblies that nucleate polymorphs. *Biomacromolecules*, 12(5):1883–1890, 2011. doi: 10.1021/bm200231c.

H. C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics*, 72(4):2384–2393, 1980. doi: 10.1063/1.439486.

D. Baker. A surprising simplicity to protein folding. *Nature*, 405:39–42, 2000. doi: 10.1038/35011000.

K. A. Ball, A. H. Phillips, D. E. Wemmer, and T. Head-Gordon. Differences in β -strand populations of monomeric A β 40 and A β 42. *Biophysical Journal*, 104(12):2714 – 2724, 2013. ISSN 0006-3495. doi: 10.1016/j.bpj.2013.04.056.

- M. N. Bannerman, R. Sargant, and L. Lue. Dynamo: a free $\mathcal{O}(n)$ general event-driven molecular dynamics simulator. *Journal of Computational Chemistry*, 32(15):3329–3338, 2011. ISSN 1096-987X. doi: 10.1002/jcc.21915.
- A. Barducci, M. Bonomi, and P. Derreumaux. Assessing the quality of the OPEP coarse-grained force field. *Journal of Chemical Theory and Computation*, 7(6):1928–1934, 2011. doi: 10.1021/ct100646f.
- R. E. Belardinelli and V. D. Pereyra. Wang-landau algorithm: A theoretical analysis of the saturation of the error. *The Journal of Chemical Physics*, 127(18):184105, 2007a. doi: <http://dx.doi.org/10.1063/1.2803061>.
- R. E. Belardinelli and V. D. Pereyra. Fast algorithm to calculate density of states. *Phys. Rev. E*, 75:046701, Apr 2007b. doi: 10.1103/PhysRevE.75.046701.
- A. M. Belcher, X. H. Wu, R. J. Christensen, P. K. Hansma, G. D. Stucky, and D. E. Morse. Control of crystal phase switching and orientation by soluble mollusc-shell proteins. *Nature*, 381(6577):56–58, 1996. doi: 10.1038/381056a0.
- A. Bellemans, J. Orban, and D. Van Belle. Molecular dynamics of rigid and non-rigid necklaces of hard discs. *Molecular Physics*, 39(3):781–782, 1980. doi: 10.1080/00268978000100671.
- A. Ben-Naim. Statistical potentials extracted from protein structures: Are these meaningful potentials? *The Journal of Chemical Physics*, 107(9):3698–3706, 1997. doi: 10.1063/1.474725.
- W. S. Bennett and T. A. Steitz. Glucose-induced conformational change in yeast hexokinase. *Proceedings of the National Academy of Sciences of the United States of America*, 75(10):4848–4852, 1978.
- T. Bereau and M. Deserno. Generic coarse-grained model for protein folding and aggregation. *The Journal of Chemical Physics*, 130(23):235106, 2009. doi: 10.1063/1.3152842.
- H. Berendsen, D. van der Spoel, and R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(13):43 – 56, 1995. ISSN 0010-4655. doi: [http://dx.doi.org/10.1016/0010-4655\(95\)00042-E](http://dx.doi.org/10.1016/0010-4655(95)00042-E).
- J. M. Berg, J. L. Tymoczko, and L. Stryer. *Biochemistry*. W. H. Freeman and Company, 5th edition, 2002.

- H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000. doi: 10.1093/nar/28.1.235.
- I. Bertini, I. C. Felli, L. Gonnelli, V. Kumar M., and R. Pierattelli. High-resolution characterization of intrinsic disorder in proteins: Expanding the suite of ^{13}C -detected NMR spectroscopy experiments to determine key observables. *ChemBioChem*, 12(15):2347–2352, 2011. ISSN 1439-7633. doi: 10.1002/cbic.201100406.
- V. Bettencourt and A. Guerra. Growth increments and biomineralization process in cephalopod statoliths. *Journal of Experimental Marine Biology and Ecology*, 248(2):191 – 205, 2000. doi: 10.1016/S0022-0981(00)00161-1.
- J. L. Bischoff. Catalysis, inhibition, and the calcite-aragonite problem; [part] 2, the vaterite-aragonite transformation. *Am. J. Sci.*, 266(2):80–90, 1968. doi: 10.2475/ajs.266.2.80.
- P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual Review of Physical Chemistry*, 53(1):291–318, 2002. doi: 10.1146/annurev.physchem.53.082301.113146.
- D. J. Brockwell, D. A. Smith, and S. E. Radford. Protein folding mechanisms: new methods and emerging ideas. *Current Opinion in Structural Biology*, 10(1):16 – 25, 2000. ISSN 0959-440X. doi: 10.1016/S0959-440X(99)00043-3.
- B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983. ISSN 1096-987X. doi: 10.1002/jcc.540040211. URL <http://dx.doi.org/10.1002/jcc.540040211>.
- A. H. Brown, P. M. Rodger, J. S. Evans, and T. R. Walsh. Equilibrium conformational ensemble of the intrinsically disordered peptide n16N: Linking subdomain structures and function in nacre. *Biomacromolecules*, 15(12):4467–4479, 2014. doi: 10.1021/bm501263s.
- C. J. Camacho and D. Thirumalai. Kinetics and thermodynamics of folding in model proteins. *Proceedings of the National Academy of Sciences*, 90(13):6369–6372, 1993.

- A. Campen, R. M. Williams, C. J. Brown, J. Meng, V. N. Uversky, and A. K. Dunker. TOP-IDP-scale: A new amino acid scale measuring propensity for intrinsic disorder. *Protein & Peptide Letters*, 15(9):956–963, 2008. doi: 10.2174/092986608785849164.
- A. Chaimovich and M. S. Shell. Relative entropy as a universal metric for multiscale errors. *Phys. Rev. E*, 81:060104, 2010. doi: 10.1103/PhysRevE.81.060104.
- M. Cheon, I. Chang, and C. K. Hall. Extending the prime model for protein aggregation to all 20 amino acids. *Proteins: Structure, Function, and Bioinformatics*, 78(14):2950–2960, 2010. ISSN 1097-0134. doi: 10.1002/prot.22817.
- M. Cheon, I. Chang, and C. K. Hall. Spontaneous formation of twisted $A\beta_{16-22}$ fibrils in large-scale molecular-dynamics simulations. *Biophysical Journal*, 101(10):2493 – 2501, 2011. ISSN 0006-3495. doi: 10.1016/j.bpj.2011.08.042.
- M. Cheon, I. Chang, and C. K. Hall. Influence of temperature on formation of perfect tau fragment fibrils using PRIME20/DMD simulations. *Protein Science*, 21(10):1514–1527, 2012. ISSN 1469-896X. doi: 10.1002/pro.2141.
- S.-W. Chiu, H. L. Scott, and E. Jakobsson. A coarse-grained model based on Morse potential for water and n-alkanes. *Journal of Chemical Theory and Computation*, 6(3):851–863, 2010. doi: 10.1021/ct900475p.
- S. Collino and J. S. Evans. Molecular specifications of a mineral modulation sequence derived from the aragonite-promoting protein n16. *Biomacromolecules*, 9(7):1909–1918, 2008. doi: 10.1021/bm8001599. PMID: 18558739.
- I. Coluzza. A coarse-grained approach to protein design: Learning from design to understand folding. *PLoS ONE*, 6(7):e20853, 2011. doi: 10.1371/journal.pone.0020853.
- T. E. Creighton. *Proteins: Structures and Molecular Properties*. Freeman, 1984.
- T. E. Creighton. *Proteins: Structures and Molecular Properties*, page 256. W. H. Freeman and Company, 2nd edition, 1992.
- L. Darré, M. R. Machado, P. D. Dans, F. E. Herrera, and S. Pantano. Another coarse grain model for aqueous solvation: WAT FOUR? *Journal of Chemical Theory and Computation*, 6(12):3793–3807, 2010. doi: 10.1021/ct100379f.
- X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, and A. E. Mark. Peptide folding: When simulation meets experiment. *Angewandte Chemie*

International Edition, 38(1-2):236–240, 1999. ISSN 1521-3773. doi: {10.1002/(SICI)1521-3773(19990115)38:1/2<236::AID-ANIE236>3.0.CO;2-M}.

- K. Delak, S. Collino, and J. S. Evans. Expected and unexpected effects of amino acid substitutions on polypeptide-directed crystal growth. *Langmuir*, 23(24):11951–11955, 2007. doi: 10.1021/la702113x.
- R. C. DeMille and V. Molinero. Coarse-grained ions without charges: Reproducing the solvation structure of NaCl in water using short-ranged potentials. *The Journal of Chemical Physics*, 131(3):034107, 2009. doi: 10.1063/1.3170982.
- R. C. DeMille, T. E. Cheatham, and V. Molinero. A coarse-grained model of DNA with explicit solvation by water and ions. *The Journal of Physical Chemistry B*, 115(1):132–142, 2011. doi: 10.1021/jp107028n.
- R. DeVane, W. Shinoda, P. B. Moore, and M. L. Klein. Transferable coarse grain nonbonded interaction model for amino acids. *Journal of Chemical Theory and Computation*, 5(8):2115–2124, 2009. doi: 10.1021/ct800441u.
- J. P. K. Doye, M. A. Miller, and D. J. Wales. Evolution of the potential energy surface with size for lennard-jones clusters. *The Journal of Chemical Physics*, 111(18):8417–8428, 1999. doi: <http://dx.doi.org/10.1063/1.480217>.
- A. Dunker, J. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic. Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling*, 19(1):26 – 59, 2001. ISSN 1093-3263. doi: 10.1016/S1093-3263(00)00138-8.
- A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradović. Intrinsic disorder and protein function. *Biochemistry*, 41(21):6573–6582, 2002. doi: 10.1021/bi012159+.
- A. K. Dunker, M. M. Babu, E. Barbar, M. Blackledge, S. E. Bondos, Z. Dosztányi, H. J. Dyson, J. Forman-Kay, M. Fuxreiter, J. Gsponer, K.-H. Han, D. T. Jones, S. Longhi, S. J. Metallo, K. Nishikawa, R. Nussinov, Z. Obradovic, R. V. Pappu, B. Rost, P. Selenko, V. Subramaniam, J. L. Sussman, P. Tompa, and V. N. Uversky. What’s in a name? Why these proteins are intrinsically disordered. *Intrinsically Disordered Proteins*, 1(1):0–4, 2013. doi: 10.4161/idp.24157.

- H. J. Dyson and P. E. Wright. Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance. In G. D. Rose, editor, *Unfolded Proteins*, volume 62 of *Advances in Protein Chemistry*, pages 311 – 340. Academic Press, 2002. doi: 10.1016/S0065-3233(02)62012-1.
- H. J. Dyson and P. E. Wright. Unfolded proteins and protein folding studied by NMR. *Chemical Reviews*, 104(8):3607–3622, 2004. doi: 10.1021/cr030403s.
- J. S. Evans. Aragonite-associated biomineralization proteins are disordered and contain interactive motifs. *Bioinformatics*, 28(24):3182–3185, 2012. doi: 10.1093/bioinformatics/bts604.
- G. Falini, S. Albeck, S. Weiner, and L. Addadi. Control of aragonite or calcite polymorphism by mollusk shell macromolecules. *Science*, 271(5245):67–69, 1996. doi: 10.1126/science.271.5245.67.
- I. C. Felli and R. Pierattelli. Novel methods based on ^{13}C detection to study intrinsically disordered proteins. *Journal of Magnetic Resonance*, 241(0):115 – 125, 2014. ISSN 1090-7807. doi: 10.1016/j.jmr.2013.10.020. A special “JMR Perspectives” issue: Foresights in Biomolecular Solution-State NMR Spectroscopy From Spin Gymnastics to Structure and Dynamics.
- A. Fernández. The principle of minimal epistemic distortion of the water matrix and its steering role in protein folding. *The Journal of Chemical Physics*, 139(8), 2013. doi: 10.1063/1.4818874.
- A. M. Ferrenberg and R. H. Swendsen. New monte carlo technique for studying phase transitions. *Phys. Rev. Lett.*, 61:2635–2638, Dec 1988. doi: 10.1103/PhysRevLett.61.2635.
- P. Frantsuzov and V. Mandelshtam. Size-temperature phase diagram for small lennard-jones clusters. *Phys. Rev. E*, 72:037102, Sep 2005. doi: 10.1103/PhysRevE.72.037102.
- D. Frenkel and B. Smit. *Understanding molecular simulation: from algorithms to applications*. Academic Pr, 2002.
- H. Fukunishi, O. Watanabe, and S. Takada. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *The Journal of Chemical Physics*, 116(20):9058–9067, 2002. doi: 10.1063/1.1472510.

- O. V. Galzitskaya, S. O. Garbuzynskiy, and M. Y. Lobanov. FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics*, 22(23): 2948–2949, 2006. doi: 10.1093/bioinformatics/btl504.
- J. Gao and D. Xu. *Correlation between posttranslational modification and intrinsic disorder in protein*, chapter 10, pages 94–103. World Scientific, 2012. doi: 10.1142/9789814366496_0010.
- S. O. Garbuzynskiy, M. Y. Lobanov, and O. V. Galzitskaya. To be folded or to be unfolded? *Protein Sci.*, 13(11):2871–2877, 2004. doi: 10.1110/ps.04881304.
- L. Gardner, D. Mills, A. Wiegand, D. Leavesley, and A. Elizur. Spatial analysis of biomineralization associated gene expression from the mantle organ of the pearl oyster *pinctada maxima*. *BMC Genomics*, 12(1):455, 2011. doi: 10.1186/1471-2164-12-455.
- S. Gomez and D. Romero. Two global methods for molecular geometry optimization. In *Proceedings of the First European Congress of Mathematics*, volume 3, pages 503–509. Birkhauser, 1994.
- Z. Guo and C. L. I. Brooks. Thermodynamics of protein folding: a statistical mechanical study of a small all-beta protein. *Biopolymers*, 42(7):745–57, 1997.
- B. He, K. Wang, Y. Liu, B. Xue, V. N. Uversky, and A. K. Dunker. Predicting intrinsic disorder in proteins: An overview. *Cell Research*, 19(8):929–949, 2009.
- G. He, T. Dahl, A. Veis, and A. George. Nucleation of apatite crystals in vitro by self-assembled dentin matrix protein 1. *Nat Mater*, 2(8):552–558, 2003. doi: 10.1038/nmat945.
- B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, 2008. doi: 10.1021/ct700301q.
- R. D. Hills, Jr, L. Lu, and G. A. Voth. Multiscale coarse-graining of the protein energy landscape. *PLoS Comput Biol*, 6(6):e1000827, 2010. doi: 10.1371/journal.pcbi.1000827.
- B. K. Ho, A. Thomas, and R. Brasseur. Revisiting the ramachandran plot: Hard-sphere repulsion, electrostatics, and H-bonding in the α -helix. *Protein Sci.*, 12(11):2508–2522, 2003. doi: 10.1110/ps.03235203.

- T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan. Geometry and symmetry presculpt the free-energy landscape of proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 101(21):7960–7964, 2004. doi: 10.1073/pnas.0402525101.
- S. A. Hollingsworth and P. A. Karplus. A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. *Biomol Concepts.*, 1(3-4): 271–283, 2010. doi: 10.1515/BMC.2010.022.
- J. D. Honeycutt and D. Thirumalai. Metastability of the folded states of globular proteins. *Proceedings of the National Academy of Sciences*, 87(9):3526–3529, 1990. doi: 10.1073/pnas.87.9.3526.
- P. Hünenberger. Thermostat algorithms for molecular dynamics simulations. In C. Dr. Holm and K. Prof. Dr. Kremer, editors, *Advanced Computer Simulation*, volume 173 of *Advances in Polymer Science*, pages 105–149. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-22058-9. doi: 10.1007/b99427.
- T. Imanaka, M. Shibazaki, and M. Takagi. A new way of enhancing the thermostability of proteases. *Nature*, 324(6098):695–697, 1986. doi: 10.1038/324695a0.
- M. R. Jensen, R. W. Ruigrok, and M. Blackledge. Describing intrinsically disordered proteins at atomic resolution by NMR. *Current Opinion in Structural Biology*, 23(3):426 – 435, 2013. ISSN 0959-440X. doi: 10.1016/j.sbi.2013.02.007.
- W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983. doi: <http://dx.doi.org/10.1063/1.445869>.
- J. Juraszek and P. G. Bolhuis. Sampling the multiple folding mechanisms of trp-cage in explicit solvent. *Proceedings of the National Academy of Sciences*, 103(43):15859–15864, 2006. doi: 10.1073/pnas.0606692103.
- L. Kalmar, D. Homola, G. Varga, and P. Tompa. Structural disorder in proteins brings order to crystal growth in biomineralization. *Bone*, 51(3):528–534, 2012. doi: 10.1016/j.bone.2012.05.009.
- F. Karush. Heterogeneity of the binding sites of bovine serum albumin. *Journal of the American Chemical Society*, 72(6):2705–2713, 1950. doi: 10.1021/ja01162a099.

- E. C. Keene, J. S. Evans, and L. A. Estroff. Matrix interactions in biomineralization: Aragonite nucleation by an intrinsically disordered nacre polypeptide, n16N, associated with a β -chitin substrate. *Crystal Growth & Design*, 10(3):1383–1389, 2010a. doi: 10.1021/cg901389v.
- E. C. Keene, J. S. Evans, and L. A. Estroff. Silk fibroin hydrogels coupled with the n16N- β -chitin complex: An in vitro organic matrix for controlling calcium carbonate mineralization. *Crystal Growth & Design*, 10(12):5169–5175, 2010b. doi: 10.1021/cg1009303.
- I. Kim, M. Darragh, C. Orme, and J. Evans. Molecular “tuning” of crystal growth by nacre-associated polypeptides. *Crystal Growth & Design*, 6(1):5–10, 2006. doi: 10.1021/cg0502183.
- I. W. Kim, E. DiMasi, and J. S. Evans. Identification of mineral modulation sequences within the nacre-associated oyster shell protein, n16. *Crystal Growth & Design*, 4(6):1113–1118, 2004a. doi: 10.1021/cg049919a.
- I. W. Kim, D. E. Morse, and J. S. Evans. Molecular characterization of the 30-AA N-terminal mineral interaction domain of the biomineralization protein AP7. *Langmuir*, 20(26):11664–11673, 2004b. doi: 10.1021/la0481400.
- J. Kim, J. E. Straub, and T. Keyes. Statistical-Temperature Monte Carlo and Molecular Dynamics Algorithms. *Phys. Rev. Lett.*, 97(5):050601, 2006. doi: 10.1103/PhysRevLett.97.050601.
- J. Kim, J. E. Straub, and T. Keyes. Statistical temperature molecular dynamics: Application to coarse-grained beta-barrel-forming protein models. *The Journal of Chemical Physics*, 126(13):135101, 2007. doi: 10.1063/1.2711812.
- J. Kim, J. E. Straub, and T. Keyes. Replica exchange statistical temperature molecular dynamics algorithm. *The Journal of Physical Chemistry B*, 116(29):8646–8653, 2012. doi: 10.1021/jp300366j.
- S. S. Kim, S. H. Kim, and Y. M. Lee. Preparation, characterization and properties of β -chitin and n-acetylated β -chitin. *Journal of Polymer Science Part B: Polymer Physics*, 34(14):2367–2374, 1996. ISSN 1099-0488. doi: 10.1002/(SICI)1099-0488(199610)34:14<2367::AID-POLB6>3.0.CO;2-T.
- J. G. Kirkwood. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3(5):300–313, 1935.

- R. Konrat. NMR contributions to structural dynamics studies of intrinsically disordered proteins. *Journal of Magnetic Resonance*, 241(0):74 – 85, 2014. ISSN 1090-7807. doi: 10.1016/j.jmr.2013.11.011. A special “JMR Perspectives” issue: Foresights in Biomolecular Solution-State NMR Spectroscopy From Spin Gymnastics to Structure and Dynamics.
- N. Kröger. The molecular basis of nacre formation. *Science*, 325(5946):1351–1352, 2009. doi: 10.1126/science.1177055.
- S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992. ISSN 1096-987X. doi: 10.1002/jcc.540130812.
- S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. Multidimensional free-energy calculations using the weighted histogram analysis method. *Journal of Computational Chemistry*, 16(11):1339–1350, 1995. ISSN 1096-987X. doi: 10.1002/jcc.540161104.
- A. Laio and F. L. Gervasio. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics*, 71(12):126601, 2008.
- A. Laio and M. Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002. doi: 10.1073/pnas.202427399.
- R. Leary. Global optima of Lennard-Jones clusters. *Journal of Global Optimization*, 11(1):35–53, 1997. ISSN 0925-5001. doi: 10.1023/A:1008276425464.
- Y.-H. Lee and B. J. Berne. Global optimization: quantum thermal annealing with path integral Monte Carlo. *The Journal of Physical Chemistry A*, 104(1):86–95, 2000. doi: 10.1021/jp991868i.
- Y. Levi, S. Albeck, A. Brack, S. Weiner, and L. Addadi. Control over aragonite crystal nucleation and growth: An in vitro study of biomineralization. *Chemistry A European Journal*, 4(3):389–396, 1998. ISSN 1521-3765. doi: 10.1002/(SICI)1521-3765(19980310)4:3<389::AID-CHEM389>3.0.CO;2-X.
- Y. Levi-Kalisman, G. Falini, L. Addadi, and S. Weiner. Structure of the nacreous organic matrix of a bivalve mollusk shell examined in the hydrated state using cryo-TEM. *Journal of Structural Biology*, 135(1):8 – 17, 2001. doi: 10.1006/jsbi.2001.4372.

- M. Lindner. libconfig, September 2012. URL <http://www.hyperrealm.com/main.php?s=libconfig>.
- P. Liu, B. Kim, R. A. Friesner, and B. J. Berne. Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13749–13754, 2005. doi: 10.1073/pnas.0506346102.
- H. Lowenstam. Minerals formed by organisms. *Science*, 211(4487):1126–1131, 1981. doi: 10.1126/science.7008198.
- H. A. Lowenstam and S. Weiner. *On Biomineralization*. Oxford University Press, 1989.
- A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wirkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B*, 102(18):3586–3616, 1998. doi: 10.1021/jp973084f.
- S. Mann. Mineralization in biological systems. In *Inorganic Elements in Biochemistry*, volume 54 of *Structure and Bonding*, pages 125–174. Springer Berlin Heidelberg, 1983. ISBN 978-3-540-12542-6. doi: 10.1007/BFb0111320.
- B. Marie, C. Joubert, A. Tayalé, I. Zanella-Cléon, C. Belliard, D. Piquemal, N. Cochenec-Laureau, F. Marin, Y. Gueguen, and C. Montagnani. Different secretory repertoires control the biomineralization processes of prism and nacre deposition of the pearl oyster shell. *Proceedings of the National Academy of Sciences*, 109(51):20986–20991, 2012. doi: 10.1073/pnas.1210552109.
- F. Marin, B. Marie, S. B. Hamada, P. Silva, N. Le Roy, N. Guichard, S. Wolf, C. Montagnani, C. Joubert, D. Piquemal, D. Saulnier, and Y. Gueguen. ‘Shel-lome’: Proteins involved in mollusc shell biomineralization -diversity, functions. In S. Watabe, K. Meayama, and H. Nagasawa, editors, *Recent Advances in Pearl Research - Proceedings of the International Symposium on Pearl Research 2011*, 2013.
- B. W. Matthews. Structural and genetic analysis of protein stability. *Annual Review of Biochemistry*, 62(1):139–160, 1993. doi: 10.1146/annurev.bi.62.070193.001035.

- H. D. Mertens and D. I. Svergun. Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *J. Struct. Biol.*, 172(1):128–141, 2010.
- N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. doi: 10.1063/1.1699114.
- R. A. Metzler, I. W. Kim, K. Delak, J. S. Evans, D. Zhou, E. Beniash, F. Wilt, M. Abrecht, J.-W. Chiou, J. Guo, S. N. Coppersmith, and P. U. P. A. Gilbert. Probing the organic-mineral interface at the molecular level in model biominerals. *Langmuir*, 24(6):2680–2687, 2008. doi: 10.1021/la7031237.
- R. A. Metzler, J. S. Evans, C. E. Killian, D. Zhou, T. H. Churchill, N. P. Appathurai, S. N. Coppersmith, and P. U. P. A. Gilbert. Nacre protein fragment templates lamellar aragonite growth. *Journal of the American Chemical Society*, 132(18):6329–6334, 2010. doi: 10.1021/ja909735y.
- M. Michenfelder, G. Fu, C. Lawrence, J. C. Weaver, B. A. Wustman, L. Taranto, J. S. Evans, and D. E. Morse. Characterization of two molluscan crystal-modulating biomineralization proteins and identification of putative mineral binding domains. *Biopolymers*, 70(4):522–533, 2003. ISSN 1097-0282. doi: 10.1002/bip.10536.
- G. Milano and F. Müller-Plathe. Mapping atomistic simulations to mesoscopic models: A systematic coarse-graining procedure for vinyl polymer chains. *The Journal of Physical Chemistry B*, 109(39):18609–18619, 2005. doi: 10.1021/jp0523571. PMID: 16853395.
- S. Miyazawa and R. L. Jernigan. Residue - residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology*, 256(3):623 – 644, 1996. ISSN 0022-2836. doi: 10.1006/jmbi.1996.0114.
- V. Molinero and W. A. Goddard. M3B: A coarse grain force field for molecular simulations of malto-oligosaccharides and their water mixtures. *The Journal of Physical Chemistry B*, 108(4):1414–1427, 2004. doi: 10.1021/jp0354752.
- V. Molinero and E. B. Moore. Water modeled as an intermediate element between carbon and silicon. *The Journal of Physical Chemistry B*, 113(13):4008–4016, 2009. doi: 10.1021/jp805227c.

- B. Monastyrskyy, A. Kryshchak, J. Moult, A. Tramontano, and K. Fidelis. Assessment of protein disorder region predictions in CASP10. *Proteins: Structure, Function, and Bioinformatics*, 82:127–137, 2014. ISSN 1097-0134. doi: 10.1002/prot.24391.
- C. Montagnani, B. Marie, F. Marin, C. Belliard, F. Riquet, A. Tayalé, I. Zanella-Cléon, E. Fleury, Y. Gueguen, D. Piquemal, and N. Cochenec-Laureau. Pmarg-pearlin is a matrix protein involved in nacre framework formation in the pearl oyster *Pinctada margaritifera*. *ChemBioChem*, 12(13):2033–2043, 2011. ISSN 1439-7633. doi: 10.1002/cbic.201100216.
- L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink. The MARTINI coarse-grained force field: Extension to proteins. *Journal of Chemical Theory and Computation*, 4(5):819–834, 2008. doi: 10.1021/ct700324x.
- J. W. Mullinax and W. G. Noid. Recovering physical potentials from a model protein databank. *Proceedings of the National Academy of Sciences*, 107(46):19867–19872, 2010. doi: 10.1073/pnas.1006428107.
- N. Nakajima, H. Nakamura, and A. Kidera. Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. *The Journal of Physical Chemistry B*, 101(5):817–824, 1997. doi: 10.1021/jp962142e.
- K. Namba. Roles of partly unfolded conformations in macromolecular self-assembly. *Genes to Cells*, 6(1):1–12, 2001. ISSN 1365-2443. doi: 10.1046/j.1365-2443.2001.00384.x.
- H. D. Nguyen, A. J. Marchut, and C. K. Hall. Solvent effects on the conformational transition of a model polyalanine peptide. *Protein Science*, 13(11):2909–2924, 2004. ISSN 1469-896X. doi: 10.1110/ps.04701304.
- C. Nogawa, H. Baba, T. Masaoka, H. Aoki, and T. Samata. Genetic structure and polymorphisms of the N16 gene in *pinctada fucata*. *Gene*, 504(1):84 – 91, 2012. ISSN 0378-1119. doi: 10.1016/j.gene.2012.03.066.
- R. Notman, E. E. Oren, C. Tamerler, M. Sarikaya, R. Samudrala, and T. R. Walsh. Solution study of engineered quartz binding peptides using replica exchange molecular dynamics. *Biomacromolecules*, 11(12):3266–3274, 2010. doi: 10.1021/bm100646z.

- C. J. Oldfield, Y. Cheng, M. S. Cortese, C. J. Brown, V. N. Uversky, and A. K. Dunker. Comparing and combining predictors of mostly disordered proteins. *Biochemistry*, 44(6):1989–2000, 2005. doi: 10.1021/bi047993o.
- E. E. Oren, C. Tamerler, D. Sahin, M. Hnilova, U. O. S. Seker, M. Sarikaya, and R. Samudrala. A novel knowledge-based approach to design inorganic-binding peptides. *Bioinformatics*, 23(21):2816–2822, 2007. doi: 10.1093/bioinformatics/btm436.
- E. E. Oren, R. Notman, I. W. Kim, J. S. Evans, T. R. Walsh, R. Samudrala, C. Tamerler, and M. Sarikaya. Probing the molecular mechanisms of quartz-binding peptides. *Langmuir*, 26(13):11003–11009, 2010. doi: 10.1021/la100049s.
- K. L. Osborne, M. Bachmann, and B. Strodel. Thermodynamic analysis of structural transitions during GNNQQNY aggregation. *Proteins: Structure, Function, and Bioinformatics*, 81(7):1141–1155, 2013. ISSN 1097-0134. doi: 10.1002/prot.24263.
- M. Ota, R. Koike, T. Amemiya, T. Tenno, P. R. Romero, H. Hiroaki, A. K. Dunker, and S. Fukuchi. An assignment of intrinsically disordered regions of proteins based on NMR structures. *Journal of Structural Biology*, 181(1):29 – 36, 2013. ISSN 1047-8477. doi: 10.1016/j.jsb.2012.10.017.
- I. Perovic, T. Mandal, and J. S. Evans. A pearl protein self-assembles to form protein complexes that amplify mineralization. *Biochemistry*, 52(33):5696–5703, 2013. doi: 10.1021/bi400808j.
- S. Piana, K. Lindorff-Larsen, and D. Shaw. How robust are protein folding simulations with respect to force field parameterization? *Biophysical Journal*, 100(9):L47 – L49, 2011. ISSN 0006-3495. doi: 10.1016/j.bpj.2011.03.051.
- J. Pillardy and L. Piela. Molecular dynamics on deformed potential energy hypersurfaces. *The Journal of Physical Chemistry*, 99(31):11805–11812, 1995. doi: 10.1021/j100031a003.
- K. W. Plaxco and M. Groβ. The importance of being unfolded. *Nature*, 386(6626):657–659, 1997. doi: 10.1038/386657a0.
- S. Pronk, S. Pll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, 2013. doi: 10.1093/bioinformatics/btt055.

- A. Rader. Coarse-grained models: getting more with less. *Current Opinion in Pharmacology*, 10(6):753–759, 2010. ISSN 1471-4892. doi: 10.1016/j.coph.2010.09.003.
- G. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7:95–99, 1963.
- G. N. Ramachandran and V. Sasisekharan. Conformation of polypeptides and proteins. *Adv Protein Chem.*, 23:283–438, 1968.
- P. Rani, A. Baruah, and P. Biswas. Does lack of secondary structure imply intrinsic disorder in proteins? A sequence analysis. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1844(10):1827 – 1834, 2014. doi: 10.1016/j.bbapap.2014.07.020.
- D. C. Rapaport. Molecular dynamics simulation of polymer chains with excluded volume. *Journal of Physics A: Mathematical and General*, 11(8):L213, 1978.
- D. Reith, M. Pütz, and F. Müller-Plathe. Deriving effective mesoscale potentials from atomistic simulations. *Journal of Computational Chemistry*, 24(13):1624–1636, 2003. ISSN 1096-987X. doi: 10.1002/jcc.10307.
- P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker. Sequence complexity of disordered protein. *Proteins: Structure, Function, and Bioinformatics*, 42(1):38–48, 2001. ISSN 1097-0134. doi: 10.1002/1097-0134(20010101)42:1<38::AID-PROT50>3.0.CO;2-3.
- B. Roux and T. Simonson. Implicit solvent models. *Biophysical Chemistry*, 78(12):1–20, 1999. ISSN 0301-4622. doi: 10.1016/S0301-4622(98)00226-9.
- T. Samata, N. Hayashi, M. Kono, K. Hasegawa, C. Horita, and S. Akera. A new matrix protein family related to the nacreous layer formation of pinctada fucata. *FEBS Letters*, 462(12):225–229, 1999. ISSN 0014-5793. doi: 10.1016/S0014-5793(99)01387-3.
- R. Samudrala and M. Levitt. Decoys ‘R’ Us: A database of incorrect conformations to improve protein structure prediction. *Protein Science*, 9(7):1399–1401, 2000. ISSN 1469-896X. doi: 10.1110/ps.9.7.1399.
- J. Seto, A. Picker, Y. Chen, A. Rao, J. S. Evans, and H. Cölfen. Nacre protein sequence compartmentalizes mineral polymorphs in solution. *Crystal Growth & Design*, 14(4):1501–1505, 2014. doi: 10.1021/cg401421h.

- C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, *The*, 27(4):623–656, 1948. doi: 10.1002/j.1538-7305.1948.tb00917.x.
- M. S. Shell. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *The Journal of Chemical Physics*, 129(14):144108, 2008. doi: 10.1063/1.2992060.
- M. R. Shirts and J. D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of Chemical Physics*, 129(12):124105, 2008. doi: 10.1063/1.2978177.
- B. A. Shoemaker, J. J. Portman, and P. G. Wolynes. Speeding molecular recognition by using the folding funnel: The fly-casting mechanism. *Proceedings of the National Academy of Sciences*, 97(16):8868–8873, 2000. doi: 10.1073/pnas.160259697.
- S. A. Showalter. *Intrinsically Disordered Proteins: Methods for Structure and Dynamics Studies*, pages 181–190. John Wiley & Sons, Ltd, 2007. ISBN 9780470034590. doi: 10.1002/9780470034590.emrstm1360.
- N. Sibille and P. Bernado. Structural characterization of intrinsically disordered proteins by the combined use of NMR and SAXS. *Biochemical Society Transactions*, 40(5):955–962, 2012. ISSN 0300-5127. doi: {10.1042/BST20120149}.
- G. N. Somero. Proteins and temperature. *Annual Review of Physiology*, 57(1): 43–68, 1995. doi: 10.1146/annurev.ph.57.030195.000355.
- Y. G. J. Sterckx, A. N. Volkov, W. F. Vranken, J. Kragelj, M. R. Jensen, L. Buts, A. Garcia-Pino, T. Jove, L. Van Melderen, M. Blackledge, N. A. J. van Nuland, and R. Loris. Small-angle X-ray scattering- and nuclear magnetic resonance-derived conformational ensemble of the highly flexible antitoxin PaaA2. *Structure*, 22(6):854–865, 2014. ISSN 0969-2126. doi: 10.1016/j.str.2014.03.012.
- Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(12):141 – 151, 1999. ISSN 0009-2614. doi: 10.1016/S0009-2614(99)01123-9.
- M. Suzuki, K. Saruwatari, T. Kogure, Y. Yamamoto, T. Nishimura, T. Kato, and H. Nagasawa. An acidic matrix protein, Pif, is a key macromolecule for nacre formation. *Science*, 325(5946):1388–1390, 2009. doi: 10.1126/science.1173793.
- R. H. Swendsen and J.-S. Wang. Replica Monte Carlo Simulation of Spin-Glasses. *Phys. Rev. Lett.*, 57(21):2607–2609, 1986. doi: 10.1103/PhysRevLett.57.2607.

- A. D. Swetnam and M. P. Allen. Improving the Wang-Landau algorithm for polymers and proteins. *Journal of Computational Chemistry*, 32(5):816–821, 2011. ISSN 1096-987X. doi: 10.1002/jcc.21660.
- W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, 76(1):637–649, 1982. doi: 10.1063/1.442716.
- J. B. Thompson, G. T. Paloczi, J. H. Kindt, M. Michenfelder, B. L. Smith, G. Stucky, D. E. Morse, and P. K. Hansma. Direct observation of the transition from calcite to aragonite growth as induced by abalone shell proteins. *Biophysical Journal*, 79(6):3307 – 3312, 2000. ISSN 0006-3495. doi: 10.1016/S0006-3495(00)76562-3.
- D. J. Tildesley. The Monte Carlo method. In M. P. Allen and D. J. Tildesley, editors, *Computer Simulation in Chemical Physics*. Kluwer Academic, 1993.
- P. Tompa. Intrinsically unstructured proteins. *Trends in Biochemical Sciences*, 27(10):527–533, 2002. ISSN 0968-0004. doi: 10.1016/S0968-0004(02)02169-2.
- P. Tompa. Intrinsically disordered proteins: a 10-year recap. *Trends in Biochemical Sciences*, 37(12):509 – 516, 2012. ISSN 0968-0004. doi: 10.1016/j.tibs.2012.08.004.
- F. Trudu, D. Donadio, and M. Parrinello. Freezing of a Lennard-Jones fluid: From nucleation to spinodal regime. *Phys. Rev. Lett.*, 97:105701, 2006. doi: 10.1103/PhysRevLett.97.105701.
- V. N. Uversky, J. R. Gillespie, and A. L. Fink. Why are ‘natively unfolded’ proteins unstructured under physiologic conditions? *Proteins: Structure, Function and Genetics*, 41(3):415–427, 2000.
- L. Verlet. Computer “experiments” on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.*, 159:98–103, Jul 1967. doi: 10.1103/PhysRev.159.98.
- M. Vihinen, E. Torkkila, and P. Riikonen. Accuracy of protein flexibility predictions. *Proteins*, 19(2):141–149, 1994.
- A. Voegler Smith and C. K. Hall. α -helix formation: Discontinuous molecular dynamics on an intermediate-resolution protein model. *Proteins: Structure, Function, and Bioinformatics*, 44(3):344–360, 2001. ISSN 1097-0134. doi: 10.1002/prot.1100.

- V. A. Wagoner, M. Cheon, I. Chang, and C. K. Hall. Computer simulation study of amyloid fibril formation by palindromic sequences in prion peptides. *Proteins: Structure, Function, and Bioinformatics*, 79(7):2132–2145, 2011. ISSN 1097-0134. doi: 10.1002/prot.23034.
- V. A. Wagoner, M. Cheon, I. Chang, and C. K. Hall. Fibrillization propensity for short designed hexapeptides predicted by computer simulation. *Journal of Molecular Biology*, 416(4):598 – 609, 2012. ISSN 0022-2836. doi: 10.1016/j.jmb.2011.12.038.
- D. J. Wales and J. P. K. Doye. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A*, 101(28):5111–5116, 1997. doi: 10.1021/jp970984n.
- S. T. R. Walsh, H. Cheng, J. W. Bryson, H. Roder, and W. F. DeGrado. Solution structure and dynamics of a de novo designed three-helix bundle protein. *Proceedings of the National Academy of Sciences*, 96(10):5486–5491, 1999. doi: 10.1073/pnas.96.10.5486.
- F. Wang and D. P. Landau. Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Phys. Rev. E*, 64(5):056101, 2001a. doi: 10.1103/PhysRevE.64.056101.
- F. Wang and D. P. Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.*, 86:2050–2053, 2001b. doi: 10.1103/PhysRevLett.86.2050.
- L. Wang, R. A. Friesner, and B. J. Berne. Replica exchange with solute scaling: A more efficient version of replica exchange with solute tempering (REST2). *The Journal of Physical Chemistry B*, 115(30):9431–9438, 2011. doi: 10.1021/jp204407d.
- Y. Wang and J. A. McCammon. Introduction to molecular dynamics: Theory and applications in biomolecular modeling. In N. V. Dokholyan, editor, *Computational Modeling of Biological Systems: From molecules to pathways*. Springer, 2012.
- S. Weiner and P. M. Dove. An overview of biomineralization processes and the problem of the vital effect. *Reviews in Mineralogy and Geochemistry*, 54(1):1–29, 2003. doi: 10.2113/0540001.

- S. Weiner, W. Traub, and S. B. Parker. Macromolecules in mollusc shells and their functions in biomineralization [and discussion]. *Phil. Trans. R. Soc. Lond. B*, 304 (1121):425–434, 1984. doi: 10.1098/rstb.1984.0036.
- M. Wells, H. Tidow, T. J. Rutherford, P. Markwick, M. R. Jensen, E. Mylonas, D. I. Svergun, M. Blackledge, and A. R. Fersht. Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proceedings of the National Academy of Sciences*, 105(15):5762–5767, 2008. doi: 10.1073/pnas.0801353105.
- D. H. Williams and I. Fleming. *Spectroscopic Methods in Organic Chemistry*, chapter 3, pages 63–72. McGraw-Hill, 5th edition, 1995.
- R. M. Williams, Z. Obradovi, V. Mathura, W. Braun, E. C. Garner, J. Young, S. Takayama, C. J. Brown, and A. K. Dunker. The protein non-folding problem: amino acid determinants of intrinsic order and disorder. In *Pacific Symposium on Biocomputing*, 2001.
- P. Wolynes, J. Onuchic, and D. Thirumalai. Navigating the folding routes. *Science*, 267(5204):1619–1620, 1995. doi: 10.1126/science.7886447.
- P. E. Wright and H. Dyson. Intrinsically unstructured proteins: Re-assessing the protein structure–function paradigm. *Journal of Molecular Biology*, 293(2):321 – 331, 1999. ISSN 0022-2836. doi: 10.1006/jmbi.1999.3110.
- H. Wu. Studies on denaturation of proteins. XIII. A theory of denaturation. *Chinese Journal of Physiology*, 5:321–344, 1931.
- B. A. Wustman, D. E. Morse, and J. S. Evans. Structural characterization of the N-terminal mineral modification domains from the molluscan crystal-modulating biomineralization proteins, AP7 and AP24. *Biopolymers*, 74(5):363–376, 2004. ISSN 1097-0282. doi: 10.1002/bip.20086.
- J. Zhang, S. M. Lewis, B. Kuhlman, and A. L. Lee. Supertertiary structure of the MAGUK core from PSD-95. *Structure*, 21(3):402–413, 2013. doi: 10.1016/j.str.2012.12.014.
- J. Zhou, I. F. Thorpe, S. Izvekov, and G. A. Voth. Coarse-grained peptide modeling using a systematic multiscale approach. *Biophysical Journal*, 92(12):4289–4303, 2007. doi: 10.1529/biophysj.106.094425.