Homologues in Semi-flexible H0P Models: A Replica-exchange Wang-Landau Study

by

Zewen Zhang

(Under the Direction of David P. Landau)

Abstract

Using replica-exchange Wang-Landau simulation, the homology of lattice proteins is checked in the semi-flexible H0P model. The short protein Crambin and five of its homologues in nature are mapped and studied in the model as examples to illustrate how this lattice protein model keeps the similarity among natural homologues. Thermodynamic and structural quantities for these lattice homologues are extracted from the simulation for comparison. From these results, we conclude that the signals in specific heat curves among these lattice homologues are similar, showing their thermodynamic similarities are retained in the H0P lattice model. But the results also cast doubt on the structural similarities of these lattice homologues. Furthermore, this study reviews the performance of the model through the investigation of degeneracy and folding signals with different values of chain stiffness.

Index words: protein folding, lattice protein model, Wang-Landau simulation, homology, degeneracy, thermodynamics, contact map

Homologues in Semi-flexible H0P Models: A Replica-exchange Wang-Landau Study

by

Zewen Zhang

A Thesis Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Master of Science

Athens, Georgia

2019

Homologues in Semi-flexible H0P Models: A Replica-exchange Wang-Landau Study

by

Zewen Zhang

Approved:

Major Professor:	David P. Landau
Committee:	Heinz-Bernd Schüttler
	Michael Bachmann

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2019

Acknowledgments

First and foremost, I would like to express my deep gratitude to Dr. David P. Landau, my research supervisor, for his patient guidance, enthusiastic encouragement and useful critiques of this research work. I would also like to thank Dr. Alfred C.K. Farris, for his advice and assistance in keeping my progress on schedule. My grateful thanks are also extended to Dr. Jiahao Xu and Dr. Zhuofei Hou for their help in programming and using computing resources, to Dr. Thomas Wüst and Dr. Guangjie Shi for providing me their codes.

Finally, I wish to thank my parents and grandparents for their emotional and financial support throughout my study.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The protein folding problem has been studied from many perspectives by biochemists, physicists, mathematicians, etc. It developed gradually into a modern interdisciplinary challenge. Huge progress has been made, while a large portion of these remarkable results come from simulation work. With the improvement of computing power and new computational algorithms [1], simulation now becomes more and more important in the study of the folding problem [2]. Among various models available for protein simulation, the coarse-grained picture, where local high-frequency motion is averaged out and atomistic structures inside an amino acid are merged into one or several "beads", is supposed to be highly efficient in computation while preserving crucial features of chemical structures and dynamics [3, 4, 5]. A widely-used example is the HP model [6, 7], which simplifies all kinds of amino acids into hydrophobic (H) or polar (P) beads restricted on sites of simple cubic lattice. This concise model proves to be very successful in globular protein systems, where it can depict the general process of protein-like folding. Using the original HP model as blueprint, some recent variants of this model take in more notable elements and therefore get more valuable results about the folding process while still holding the remarkable efficiency during simulation. In our previous research, two enriched lattice protein models [8, 9] are proposed to overcome the

high-degeneracy problem in the HP model by introducing an extra "neutral" (0) monomer and rigidity within the polymer chain, respectively. These two models significantly reduce ground degeneracies for lattice proteins, and then merge into the semi-flexible H0P model to further bring together their advantages. Simulation of this new model is conducted with the help of the Wang-Landau sampling [10] method and its enhanced parallel version [11]. The new model enables measurements of degeneracy [12], density of sates [12], structural properties [13], contact maps [14], etc. As a result, our recent study with the semi-flexible H0P model offers robust degeneracy in low-lying state structures and folding behaviors that are lucid and easily accessible. In this stage, it is worthwhile to apply the model to a wider range of sequences.

On the other hand, the tremendous number of different proteins and their mysterious dynamics hinders the numeration and conclusion of protein folding problems. But progress has been achieved in sorting protein sequences and folds [15], which casts light onto some deeper understanding to the problem. Therefore, searching for possible similarity among complex folding behaviors becomes an important method to decode the puzzle of protein folding [16, 17]. Although in general cases, natural homologue sequences tend to possess similar folds [18, 19, 20, 21], more and more experiments indicate that slight changes in protein sequences can lead to catastrophic differences in secondary [22] /tertiary [23] structures and stability [24, 25, 26]. Same results are also observed in simulation with various models, such as mutation in the aforementioned HP model [27, 28, 29] and other lattice protein systems [30, 31]. These simulations enable us to understand the folding behaviors of similar sequences and extract the common features of their folding pathway. Meanwhile, testing the similar sequences serves good opportunities to review those models applied in simulation and figure out how far a model deviates from its counterpart in nature. For these purposes, natural homologues become good objects to study with simulation.

As a widely used example for simulation, the short protein, Crambin, has been mapped in different lattice models and studied. Besides its simple sequence, the advantage of testing Crambin is its precisely measured structure which enables delicate comparison between experiments and simulation. Thus, in this thesis, Crambin is applied as a prototype for simulation. Furthermore, five homologues of Crambin in nature [32] are mapped into the H0P model for simulation. Previous research indicated the existence of similarity on thermodynamic evolution of these lattice homologues, but questions still remain. In this study, these proteins are analyzed for both statistical results and structural details. In Chapter 2 and Chapter 3, we describe the models and methods applied in this research respectively. Results are presented in Chapter 4 and Chapter 5. Further discussion about the model can be found in Chapter 6, and conclusion is drawn in Chapter 7.

Chapter 2

Model

2.1 Protein Folding Problem

2.1.1 Amino Acids and Peptide Chains

Proteins are usually the largest component in living cells other than water, performing the major biochemical reactions and serving all kinds of life processes. A protein is a biomolecule with long-chain structure (so called *polypeptide chain*), made up of merely 20 kinds of amino acids. As shown in Fig. 2.1, the structure of an amino acid is organized around a central carbon atom, namely C_α. Connected to the C_α, there are a hydrogen atom, an amino group, a carboxyl group and a functional group R. Amino acids differ from each other only in their side chain R, which leads to the diversity in their hydrophobicity and therefore provides the criteria of monomer classification in the HP model. The amino groups and carboxyl groups in two amino acids are ionized and rejointed to form a peptide bond. When a series of amino acids is connected by peptide bonds in a specific order, they construct a polypeptide chain, which is commonly known as the *primary structure* of protein. The chain of C_α is usually called the "backbone", while the repeating units along the backbone are referred as "residues".

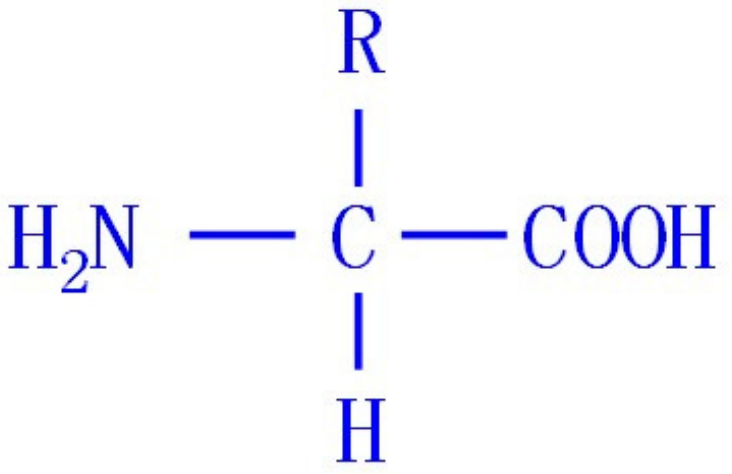


Figure 2.1: The structure of amino acids

In a proper aqueous environment, the polypeptide chain may fold into its specific conformational structure, and then get ready to serve certain biological functions. These folded structures, mainly curling into α-helices or constraining into β-sheets due to the combination of hydrogen bonds (NH groups and C=O groups are donors and acceptors, respectively) , are called *secondary structures*. Those helices and sheets may furthermore arrange in irregular architectures called *tertiary structures*. However, if the environment changes, a protein will deviate from its native structure. If temperature increases, a folded protein will melt into a globule state and finally unfold into some random coils. If pH value changes, a folded protein will also denature. However, upon restoring the original environment, a denatured protein may return to its native structure. The time scale for the restoring process varies only from seconds to minutes. The question why the restoring process is so fast and cooperative leads to the proposal of folding pathway theory [33].

There are also quaternary structures existing in some protein systems, which are formed by the assembly of multiple proteins. But tertiary structures are currently the main target for current protein simulation research to predict.

On the other hand, due to the interactions among residues and medium, which are mainly hydrogen bonds and occasionally disulfide bonds, the free energy landscape for a protein can be very complex with many local minima. Therefore, there is a famous "Levinthal paradox" in the kinetic part of protein folding, which indicates that it is unlikely for a protein to randomly search the landscape for its native state in limited time. The fact of two stages in the protein folding process (denatured state $\rightarrow$ molten globule $\rightarrow$ native state) suggests the existence of folding "funnels", and a folding "pathway" in the funnel to guide the rapid folding process.

2.1.2 Coarse-grained Models

The length of proteins usually varies from 50 to 3000 residues, while some huge proteins contain more than 20000 residues. Thus, it can be inferred from an easy estimation that the number of proteins in nature is really tremendous ($>20^{3000}$). However, only hundreds of thousands of proteins are measured at $\mathring{A}$-level resolution [34], even though incredible progress has been made in experimental techniques. These are only a small fraction of the known protein sequences. Therefore, to draw the complete map of protein folding, predicting native structures of proteins remains a challenge for theoretical and computational research for the foreseeable future. In this stage, simulation of the protein folding process is supposed to serve as powerful tools for understanding the mechanism of protein functions.

In the classical way, there are two different perspectives to simulate a protein. The first is all-atom modeling, which practically keeps every individual atom during the simulation. This modeling does molecular dynamics (MD) or Monte Carlo (MC) simulation at the atomistic scale. Limited by modern computing resources, this bottom-up approach can

be only applied to very short proteins for a rather fast folding process. So, the application of all-atom simulation is restricted to small systems in many cases and different levels of details in simulation models are proposed for improvement. Therefore, the second choice is to implement coarse-grained models. From the coarse-grained perspective, the number of variables in a model is supposed to be minimized properly to accelerate simulations. To do that, atomistic details are averaged out in mesoscopic scales with respect to the experimental data, so that functional groups, residues or covalent bonds in the polypeptide chain are treated using simple geometries, such as balls or sticks. This simplification enables the application to larger-scale protein systems for longer physical time [35] in the simulation. Though possible problems may come along with the low resolution in coarse-grained models, these models are capable of providing important evidence for folding behaviors [35], including protein structure prediction [36], environmental effects [37, 38] and dynamics [39]. More specifically, there are different levels of coarse views to reduce the complexity of polypeptide chains, *e.g.* using single or multiple parts [40] to represent a residue, applying round-bead or anisotropic shapes [41] to simplify the structures of monomers, and placing polymers on on-lattice or contineous off-lattice frame [42]. Meanwhile, interaction between monomers in coarse-grained models may also derived via "*physics-based*" [43, 44] or "*knowledge-based*" [45, 46] methods.

2.1.3 Homologues

The term *homologue* was first introduced to describe the evolutionary relation between two macroscopic biological entities. When this concept is generalized to the field protein research, it points to some proteins sharing a common ancestor. The so-called *homology* can be determined by the comparison of sequence, structural and functional similarity among different proteins. If there are enough physicochemical evidences to support the same evolutionary root of two proteins, these two proteins are identified as homologues, or they are believed to have "homology". However, as *homology* is a common term in the protein research field, sometimes it generally refers to the similarity of sequences in literature [47].

Homology plays one of the most vital roles in structural biology. In nature, it is safe to predict that two proteins with over 40% similarity in their sequences tend to fold into the same structure [48], and thus preform similar functions. However, according to many previous experimental results, the folds of some kinds of highly similar sequences may still be different [49, 50]. There is also a term *divergent evolution* in protein study, which means two homologues have similar sequences but different mechanisms. Therefore, the question how the native structures and functions of proteins are related to their sequence remains even for homologues.

For protein simulation, homology is also very important. It is applied to guide the foundation of new modeling methods [51] and provide templates or databases for structure prediction [52]. Therefore, it is worthy to test the homology when a simulation model is proposed, especially to start with the simulation of natural homologues in the model. Examination on homologues will provide us information about how the model deviate from its nature counterpart. For example, if two sequences with a certain percentage of mismatches between their sequences have similar thermodynamical, structural and functional properties, will they hold these similarities in the simulation model?

There are some families of homologues whose native structures are already precisely measured, which may serve the comparison of simulation and experiments. One famous example is the small plant toxin, Crambin, which has for long been used in different simulations, and its homologues. In this thesis, the short protein, Crambin, as well as five of its homologues (PDB entry *1bhp, 1wuw, 1jmn, 1orl, 3c8p*) in nature [32], are mapped and studied.

2.2 Semi-flexible H0P Lattice Model

2.2.1 HP Model

For globular protein systems in aqueous environment, the dominant driving force for folding is hydrophobic effect, which mainly arises from the formation and competition of hydrogen bonds among residues and solvent molecules. The polar molecules in the solvent seek to pair with polar (hydrophilic) residues, and push nonpolar (hydrophobic) residues away. To minimize the contact energy, nonpolar residues are squeezed into the middle of folded structures. Thus, the native structure of a globular protein usually looks like a solid hydrophobic ball covered with a polar shell.

In consideration of the hydrophobic effect, the HP model [6] was proposed. Like other coarse-grained models, the HP model treats residues in a polypeptide chain as single beads connected with inelastic sticks (rigid covalent bonds), and considers interactions in the polymer system regardless of shapes of residues. The HP model condenses 20 kinds of amino acids into 2 categories: hydrophobic (H) and polar (P), and put those monomers on the site of simple cubic lattice. Then a self-avoiding walk is performed in two- or three-dimensional lattices to investigate the properties of the lattice protein system. For simplicity, only contacts between two non-bonded nearest (one lattice constant) neighbors in the structure are taken into consideration, while energy assigned to those contacts is only determined by categories of two paired monomers (H-H, H-P and P-P). Most HP models take H-H contacts

as the only energetic term, which give the Hamiltonian written as:

$$\mathcal{H} = -\epsilon_{HH} n_{HH}, \tag{2.1}$$

where n_{HH} indicates the number of H-H contacts and ϵ_{HH} gives the H-H contact energy. Besides the success of the HP model in simulational and mathematical [53, 54] perspectives, this binary pattern is also supported by experiments [55].

2.2.2 H0P Model

Previous research has indicated that the high degeneracy in native states hinders the HP model to thoroughly explore the conformational space [12]. To reduce the degeneracy, a third kind of "neutral" monomer (0), whose hydrophobicity is between the H and P monomers, is introduced into the model [9]. Therefore, the new model is named H0P model. Analogous to the HP model, the H0P model only considers contact energy between two non-bonded nearest neighbors, which is determined by categories of two paired monomers. Commonly, only H-H and H-0 contacts are taken into consideration. In this way, the Hamiltonian for the model is given by:

$$\mathcal{H} = -\epsilon_{HH} n_{HH} - \epsilon_{H0} n_{H0}, \tag{2.2}$$

where n_{HH} indicates the number of H-H contacts and n_{H0} indicates the number of H-0 contacts; ϵ_{HH} gives the H-H contact energy and ϵ_{H0} gives the H-0 contact energy.

2.2.3 Semi-flexible H0P Model

In spite of its huge improvement to the original HP model, the H0P model may suffer from considerably high degeneracy in native state as well [8]. Therefore, in view of the effect of internal rigidity within natural polypeptide chains [56], the concept of "stiffness" is

introduced into the model [8, 57]. This enriched model assigns an additional energetic term to each "bend" in a lattice protein structure, which represents the positive contribution to internal energy from any bends in structures. Therefore, the Hamiltonian of the semi-flexible H0P lattice model is given by:

$$\mathcal{H} = -\epsilon_{HH} n_{HH} - \epsilon_{H0} n_{H0} - \epsilon_{\theta} n_{\theta}, \tag{2.3}$$

where n_{HH} is the number of H-H contacts, n_{H0} is the number of H-0 contacts and n_{θ} is the number of bends; ϵ_{HH} is the H-H contact energy, ϵ_{H0} is the H-0 contact energy and ϵ_{θ} is the bending energy. The energy for all other kinds of contacts are ignored in the model for clarity. The value of ϵ_{θ} is usually set to negative to intuitively show the fact that polymer stiffness is a competitive effect against the formation of contacts with the decrease of temperature. Fig. 2.2 illustrates the Hamiltonian shown in Eq. 2.3.

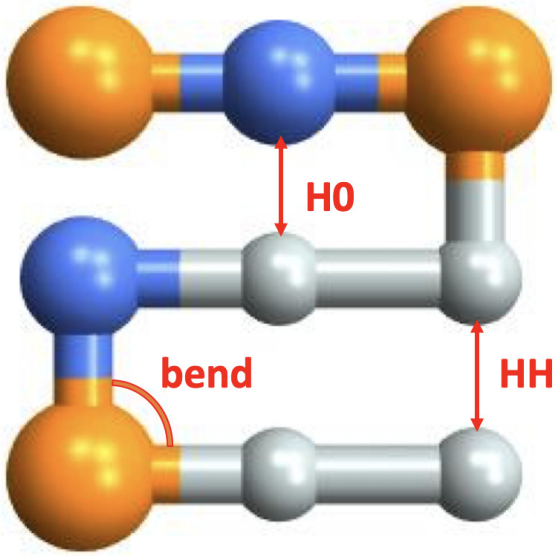


Figure 2.2: An example of semi-flexible H0P lattice protein
Small white beads represent H monomers; medium blue beads represent 0 monomers; large orange beads represent P monomers. In this demonstrative structure, there are one H-0 contact, two H-H contacts and four bends that contribute to the internal energy under the frame of the semi-flexible H0P model.

The parameters in this project are chosen as ($\epsilon_\theta = -0.1\epsilon_{H0} = -0.05\epsilon_{HH}$), which are proved to be appropriate for simulating lattice Crambin [8] with clear transitions and low ground state degeneracy.

2.2.4 Mapping for Crambin Homologues

Crambin has been mapped and studied in lattice protein models [8, 12, 58]. Meanwhile, sequences and structures of Crambin itself and its homologues (small plant toxins, such as purothionin and viscotoxin) have long been known by experiment at high resolution [32, 59], as shown in Fig. 2.3. Therefore, these small proteins are an ideal starting point for us to look into similarities among homologues in the semi-flexible H0P model.

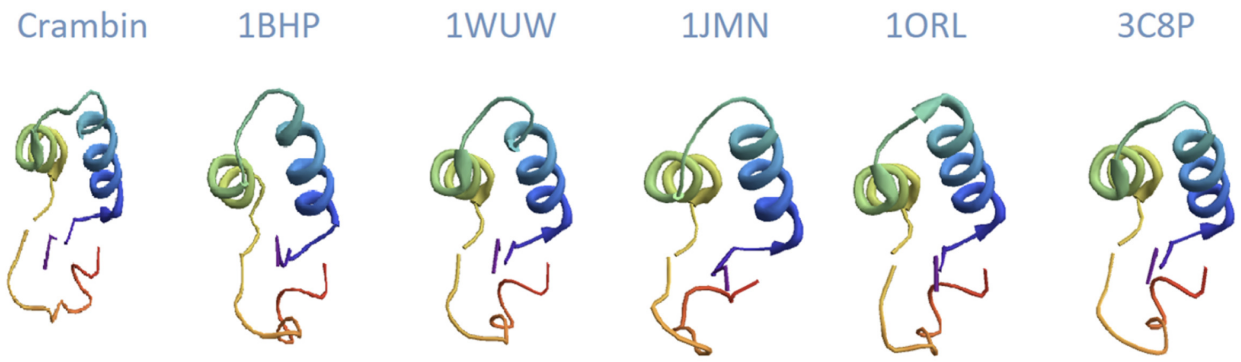


Figure 2.3: Native structures of Crambin and 5 homologues. (Source: PDB)

In this thesis, one purothionin (PDB entry *1bhp*), one hordothionin (*1wuw*) and three viscotoxins (*1orl*, *1jmn* and *3c8p*) are chosen to be tested as Crambin homologues. Their sequences are mapped into an H0P form, as shown in Table 2.1. Although there are only three kinds of monomers in the H0P representation, these homologues still have significant differences in their sequences. However, comparing to their differences in real sequences, which are around 50%, they are more "alike" in the H0P model.

The percentages of H, 0 and P monomers in these homologues are also different, as concluded in Table 2.2. The distribution of different kinds of monomers in these converted

Protein	**Sequence in H0P Model**									
Crambin	PPHH0	0HHHP	0PHPH	HPH00	00P0H	H0000	0HHHH	0000H	00P00	P
1bhp	P0HHP	00H0P	PH0PH	HP0P0	0PP-H	H0PHH	PHPH0	00H0H	0PPH0	P
1wuw	P0HHP	00H0P	PH0PH	HPHP0	0PP-H	H0P0H	PHPH0	00HPH	000H0	P
1orl	P0HH0	P000P	PH0P0	HPH00	00PPP	H0PH0	0HPHH	0000H	00P00	P
1jmn	P0HH0	P000P	PH0P0	HPH00	00PPH	H00H0	0HPHH	0000H	00P00	P
3c8p	P0HH0	0000P	PH0P0	HPH00	00PP0	H0PH0	0HPHH	0000H	00P00	P

Table 2.1: Sequences of Crambin homologues in the H0P model

The dash (-) here represents for one missing monomer, which helps to align the sequences.

	Crambin	*1bhp*	*1wuw*	*1orl*	*1jmn*	*3c8p*
H(%)	34.8	33.3	33.3	23.9	26.1	23.9
0(%)	45.7	33.3	35.5	47.8	50.0	52.2
P(%)	19.6	33.3	31.1	28.3	23.9	23.9

Table 2.2: Percentages of H, 0 and P monomers in Crambin homologues

sequences is rather balanced.

Higher percentage of H and 0 monomers in an H0P sequence usually leads to wider internal energy range.

Chapter 3

Simulation Methods

3.1 Thermodynamics and structural quantities

In statistical mechanics, the partition function ($Z(T)$) at temperature T can be transformed into a summation over the energy space when the density of states $g(E)$ is known for the system of interest:

$$Z(T) = \sum_{\{E_i\}} g(E_i) \exp(-E_i/k_B T), \tag{3.1}$$

where $g(E_i)$ is the temperature-independent density of states at a certain energy E_i, and k_B is the Boltzmann constant. Once $g(E_i)$ is measured, other average quantities can be calculated from

$$\langle Q(T) \rangle = \frac{\sum_{\{E_i\}} \overline{Q}(E_i) g(E_i) \exp(-E_i/k_B T)}{Z(T)}, \tag{3.2}$$

where $\overline{Q}(E_i)$ is the average quantity of interest Q at energy E_i.

The heat capacity C_V can then be derived from

$$C_V(T) = \frac{\langle E^2 \rangle - \langle E \rangle^2}{k_B T^2}, \tag{3.3}$$

where $\langle E^2 \rangle$ and $\langle E \rangle$ follow the calculation in Eq. 3.2.

Another commonly used structural quantities in protein research is the radius of gyration (r_g), given by

$$r_g = (\frac{1}{N}\sum_{i=1}^{N}(\vec{r}_i - \vec{r}_{CM})^2)^{1/2}, \tag{3.4}$$

where N is the number of monomers in the protein, $\vec{r}_i$ is the position of the i^{th} monomer and $\vec{r}_{CM}$ is the position of the center of mass.

3.2 Wang-Landau Simulation

3.2.1 Wang-Landau Simulation

Monte Carlo (MC) methods refer to a class of computational or simulational methods that implement massive random sampling to obtain results. The most famous method in MC is Metropolis sampling [60], which does random work in configuration space. However, for systems with complex free energy landscape, classical MC methods like Metropolis sampling can easily get stuck in local minima at low temperature.

Wang-Landau sampling overcomes this drawback by performing random walk in energy space, so that it can estimate the density of states ($g(E)$) for a system. During the MC process, a histogram ($H(E)$) is kept to make sure that every energy state is evenly visited. Meanwhile, the temperature-independent acceptance probability of one MC trial relating to the current configuration A and a proposed configuration B is given by

$$P(A \to B) = min\{1, \frac{g'(E_A)}{g'(E_B)}\}, \tag{3.5}$$

where $g'(E_A)$ and $g'(E_B)$ are current estimators of the $g(E_A)$ and $g(E_B)$, respectively. After

a trial move, the histogram will be updated by,

$$H(E_i) = H(E_i) + 1 \tag{3.6}$$

where E_i is the energy of the accepted or remained state. The $g'(E_i)$ is updated as well according to

$$g'(E_i) \to f \times g'(E_i) \tag{3.7}$$

where f is the modification factor with initial value larger than 1 (*e.g.* $f = e$). For a certain number of trial moves, the histogram is checked for flatness; if satisfying a preset criterion (*e.g.* 80%), all bins in the histogram $H(E)$ will be reset to zero, and the value of f will be lowered, such as $f \to \sqrt{f}$. Then the next round of iteration begins with the current value of $g'(E)$. The simulation is finished when f is small enough, and then provides its final estimation of $g(E)$.

3.2.2 Replica-exchange Wang-Landau Simulation

For systems with a wide free energy range, sometimes the regular Wang-Landau algorithm can take too much time to search the entire energy space. Thus, the replica-exchange Wang-Landau algorithm (REWL) is introduced to enhance the efficiency of Wang-Landau sampling by taking advantage of parallel computing [11]. It divides the whole energy range into multiple overlapping energy windows; inside each window, there are one or more random walkers doing Wang-Landau sampling independently on different configurations (replicas). Replica-exchanges, a technique that attempts to swap the configurations of two random walkers to expedite the sampling, are proposed between two walkers in the overlapping area of two neighboring windows i and j regularly at fixed interval, with the acceptance

probability given as:

$$P_{switch} = min\{1, \frac{g_i'(E_A)}{g_i'(E_B)} \frac{g_j'(E_B)}{g_j'(E_A)}\}, \tag{3.8}$$

where $g_k'(E_X)$ represents the estimator for the density of states from walker k in energy level X. When simulation is finished, the estimators in all energy windows can merge properly to get the final estimated density of states. In this way, REWL allows us to efficiently search for complex energy landscapes.

3.2.3 2D Wang-Landau Simulation

Wang-Landau sampling can also be generalized into higher dimensions, taking in other variables like some structural quantities. In 2D cases, random walkers do sampling in a 2D space expanded by energy E and another quantity Q, calibrating the 2D density of states, $g(E, Q)$. Now the acceptance of a proposed MC move becomes

$$P(A \to B) = min\{1, \frac{g'(E_A, Q_A)}{g'(E_B, Q_B)}\}, \tag{3.9}$$

where $g'(E_A, Q_A)$ and $g'(E_B, Q_B)$ are estimators of the $g(E_A, Q_A)$ and $g(E_B, Q_B)$, respectively. The remaining procedures like updating the histogram, estimators and modification factor are the same as those in the 1D Wang-Landau method. The 2D Wang-Landau method can also take in the replica-exchange techniques, that is, dividing the 2D variable space into overlapping rectangular windows, performing sampling inside each window and proposing replica-exchanges between nearby windows.

3.2.4 Replica-exchange Multicanonical Sampling

When the final estimation of $g(E)$ is extracted from the system of interest, other quantities can be measured with replica-exchange multicanonical sampling (REMUCA). During the

REMUCA simulation, the known $g(E)$ is kept constant, and applied to the acceptance probability in Eq. 3.5. The proposal and acceptance of replica exchange is still the same as Eq. 3.8. As there is no modification factor f to help terminate the simulation, a preset number of Monte Carlo steps is required here.

During the simulation, a two-dimensional histogram $H(E, Q)$ for quantities (Q) other than the internal energy is used to trace the frequency of Q. It is updated as $H_i(E, Q) \rightarrow H_i(E, Q) + 1$ after each MC step. When the REMUCA simulation is finished, the jointed density of states $g(E, Q)$ is obtained by $g_i(E, Q) = g_i(E)H_i(E, Q)$.

3.3 Trial Moves

To satisfy the requirement of both ergodicity and efficiency, three types of trial moves for lattice proteins are used: pivot moves (Fig. 3.1), pull moves [61] (Fig. 3.2) and bond rebridging [62] (Fig. 3.3).

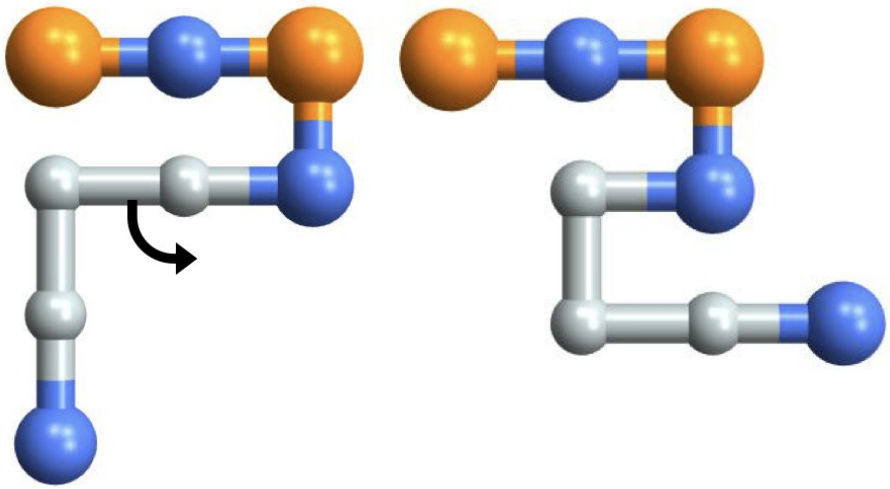


Figure 3.1: Pivot move
A monomer in the middle is chosen as "pivot". Part of the polymer rotates around the "pivot" by 90° to get the new configuration on the right.

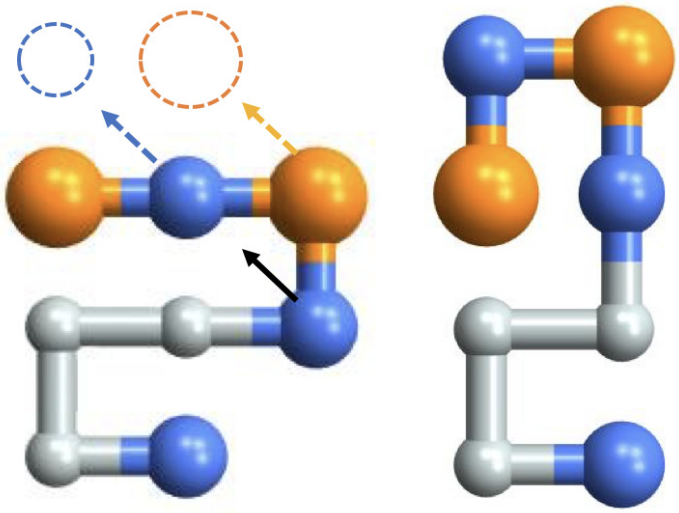


Figure 3.2: Pull move
A monomer in the middle moves to a nearby position. To keep a complete "single chain" structure, part of the polymer is "pulled" to new positions. This gives the new configuration on the right.

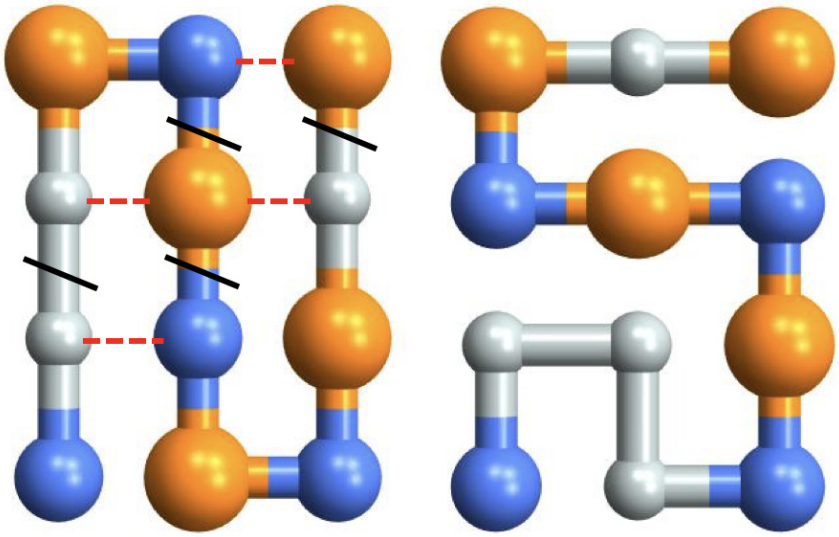


Figure 3.3: Bond rebridging
A pair of bonds are cut and reconnected in its orthogonal direction. This gives the new configuration on the right. The monomers needs to be reassigned to each site in the new structure according to their order in the sequence.

Chapter 4

Results: Thermodynamic Properties

4.1 Density of States

4.1.1 Density of States

With the help of the Wang-Landau sampling methods, the first result we can get is the density of states for all these lattice proteins. The density of states for these lattice homologues are plotted along with the results for Crambin in Fig. 4.1 and Fig. 4.2. Generally, the shapes of density of states for all tested homologues are similar. The major difference is energy ranges that different lattice homologues span. In the H0P mapping, there are differences in the percentage of "H", "0" and "P" monomers (see Table 2.2) among all tested sequences, which mainly lead to changes in the internal energy ranges. For lattice *1bhp* and *1wuw*, although their sequences have 5 different monomers (11%) in the H0P representation, their specific heat curves are still highly similar. Same results can be noticed when bringing together the specific heat curves of lattice *1orl* and *3c8p*.

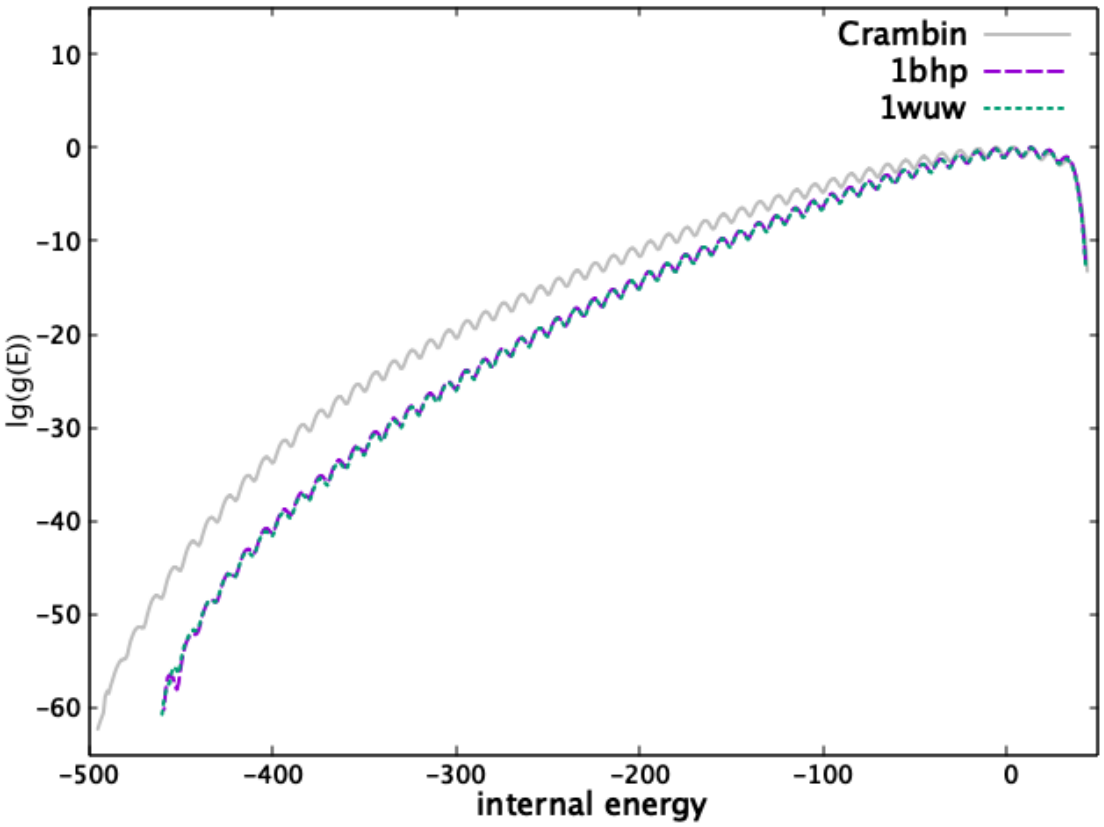


Figure 4.1: Density of states for H0P lattice homologues *1bhp* and *1wuw* plotted along with the results for Crambin
The error bars are smaller than the width of the curves. The sawtooth shape of the curves are caused by the rearrangement of bends when the number of contacts is kept. The curves for lattice *1bhp* and *1wuw* are plotted together because their sequences are highly alike.

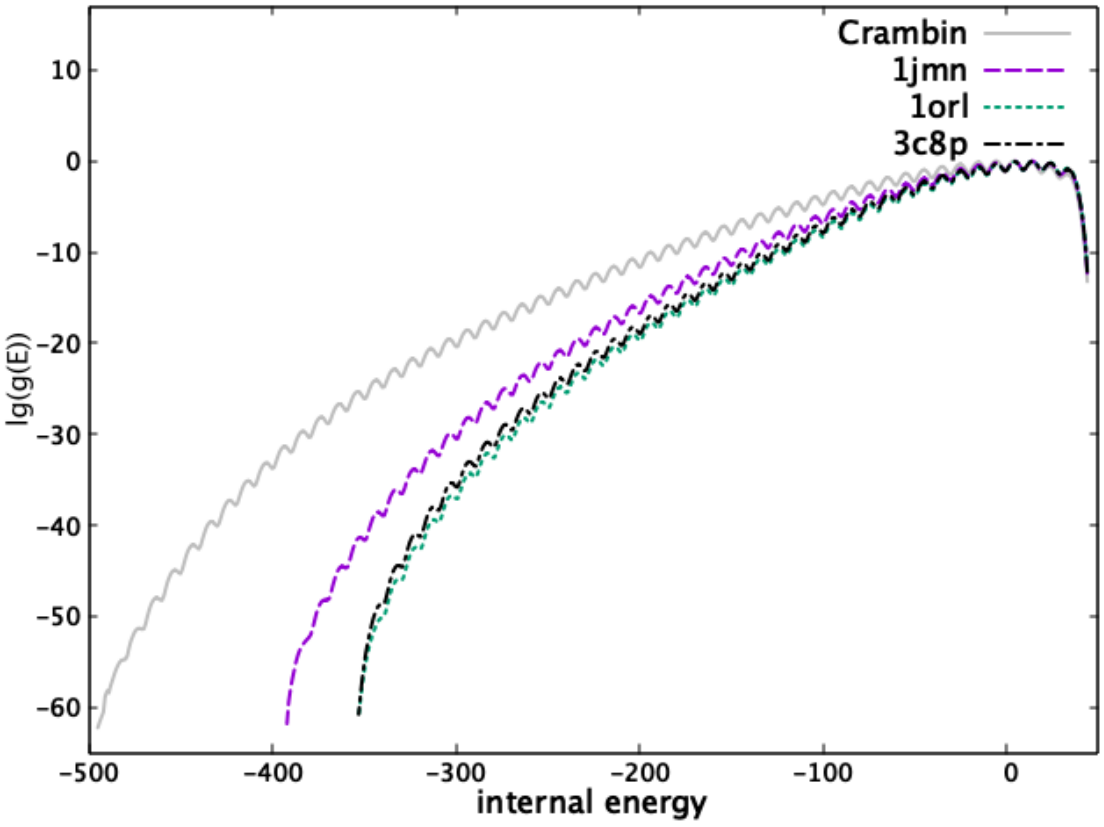


Figure 4.2: Density of states for H0P lattice homologues *1jmn*, *1orl* and *3c8p* plotted along with the results for Crambin

The error bars are smaller than the width of the curves. The curves for lattice *1jmn*, *1orl* and *3c8p* are plotted together because they all belong to the viscotoxin family.

4.1.2 Ground State Degeneracy

Searching for the native state structures is an important part of a lattice protein model. When REWL simulation gives the lowest energy level, REMUCA is applied to search for all possible native structures. As listed in Table 4.1, the degeneracies of all native states turn out to be low, which supports the belief that the parameters chosen in **2.2.3** are suitable for simulating this family of lattice proteins. To a certain extent, the real degeneracies for some lattice homologues may be lower than the numbers listed in Table 4.1, as one slight shift of some non-bonded monomers can change a native structure to a different one, while the two structures has no nontrivial differences physically (see Fig. 4.3 for example). However, some lattice homologues (*e.g. 3c8p*) still show relatively high degeneracy in their low-lying excited states.

Energy level	Crambin	*1bhp*	*1wuw*	*1orl*	*1jmn*	*3c8p*
g(E_0)	1	3	2	3	2	5
g(E_1)	6	26	10	24	24	47
g(E_2)	0	84	17	133	118	362
g(E_3)	6	119	24	633	358	1810

Table 4.1: Degeneracy of ground states and low excited states

g(E_i) represents the density of states for the i^{th} excited state from the ground state. E_0 is the native state. 0 in g(E_2) of Crambin indicates no state has been found at this energy level, which might be a gap in the energy spectrum.

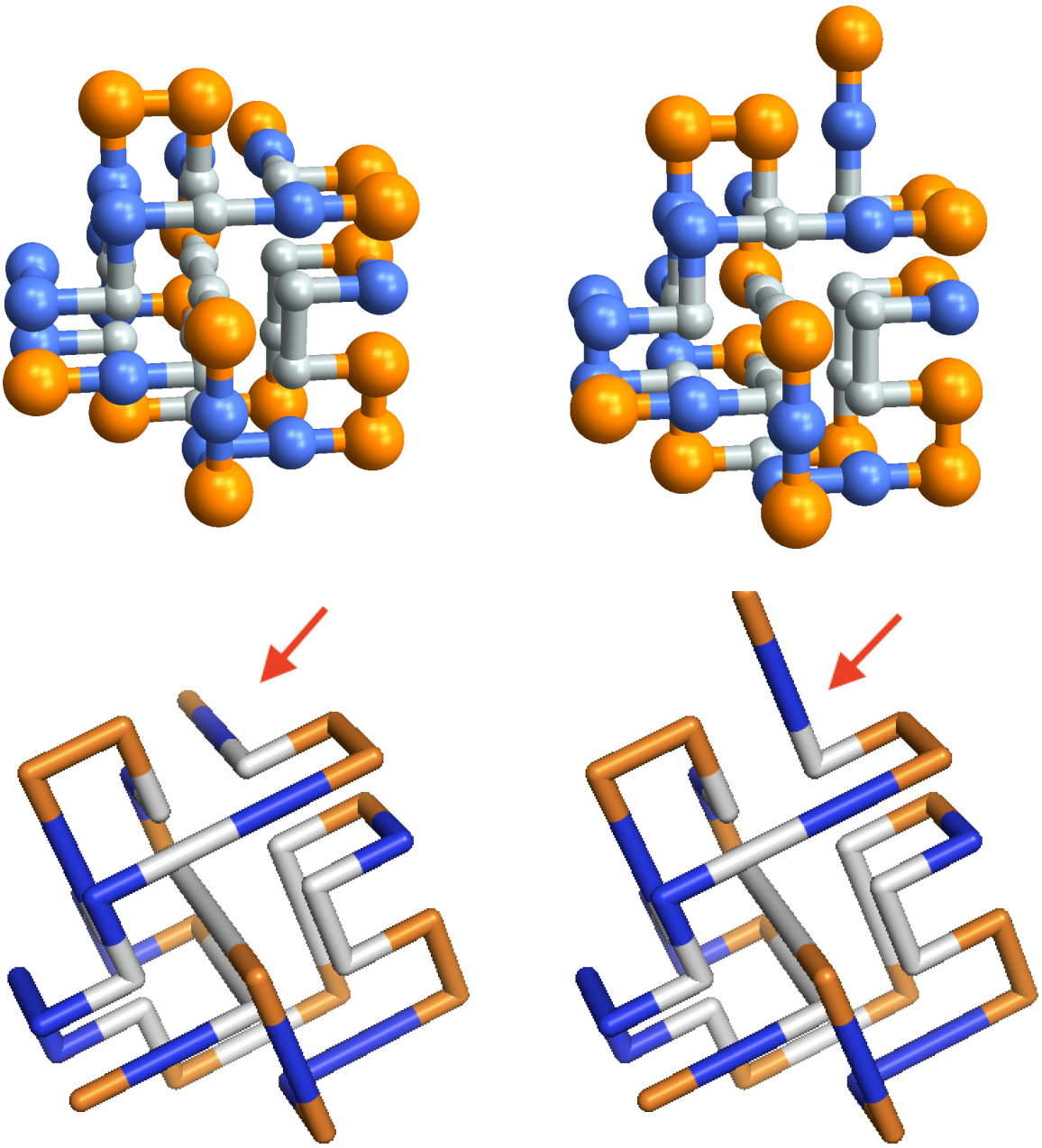


Figure 4.3: Two native structures of H0P lattice protein *1bhp*
For clarity, the lower two graphs use "sticks" instead of "beads" to represent monomers, so that those obstructed monomers in the back can be visualized. As pointed by arrows, the only difference between these two structures is the shift of two monomers in the end of the polymer.

4.2 Specific Heat

4.2.1 Specific Heat Curve

After the density of sates is properly extracted from simulation, internal energy can be derived from Eq. 3.2 and specific heat (C_V) can also be calculated according to Eq. 3.3. The results are shown in Fig. 4.4 and 4.5. For all these lattice homologues, their specific heat curves show two signals in the relatively high temperature region, corresponding to the folding transition and coil-globule transition, although these transitions may happen at slightly different temperature in this model. Furthermore, in the low temperature region, there are also two signals observed at fixed temperature (T = 0.02 and T = 0.05), which indicate that the native states of all these homologues start "melting" at the same temperature. Therefore, these lattice homologues show similar thermodynamic behaviors in this model.

4.2.2 Transition

As mentioned in **4.2.1**, there are two signals at high temperature, corresponding to the folding transition and coil-globule transition. Three typical structures before and after these two transitions are shown in Fig. 4.6 and Fig. 4.7.

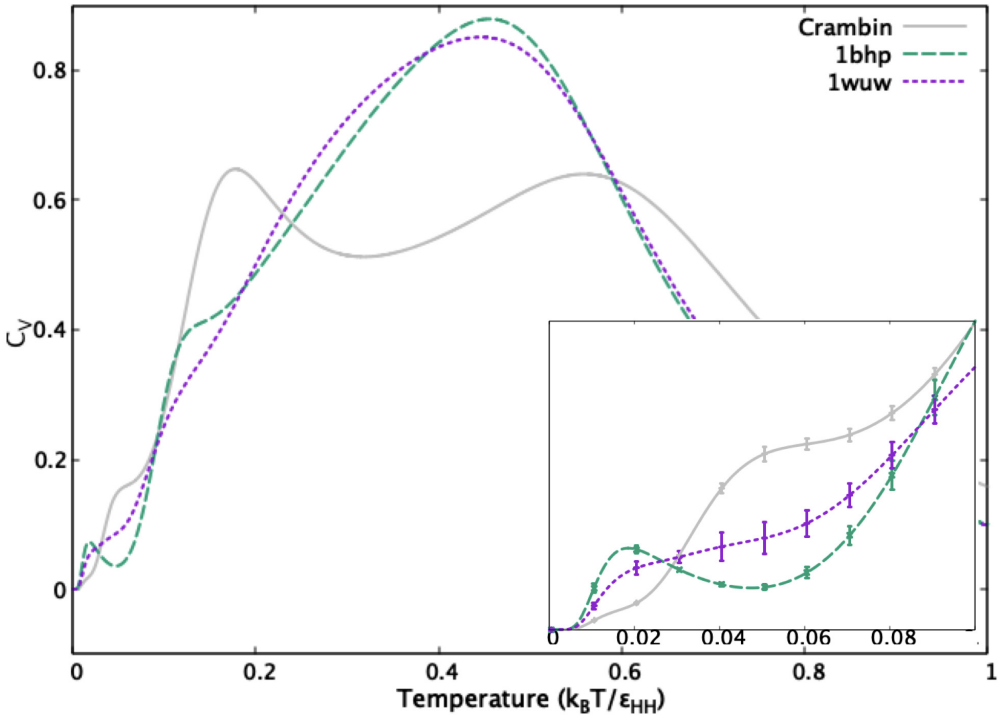


Figure 4.4: Specific Heat curves for two H0P lattice homologues *1bhp* and *1wuw* along with the curve for lattice Crambin

Part of the curves at low temperature region is magnified for clarity. Error bars smaller than the width of lines are ignored in the main graph.

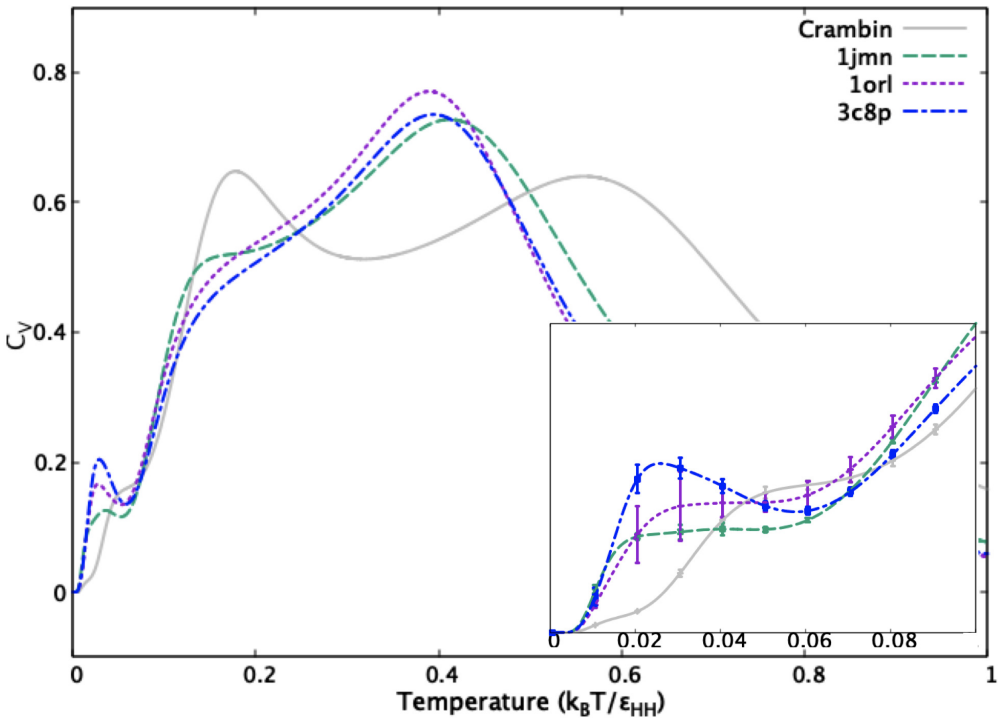
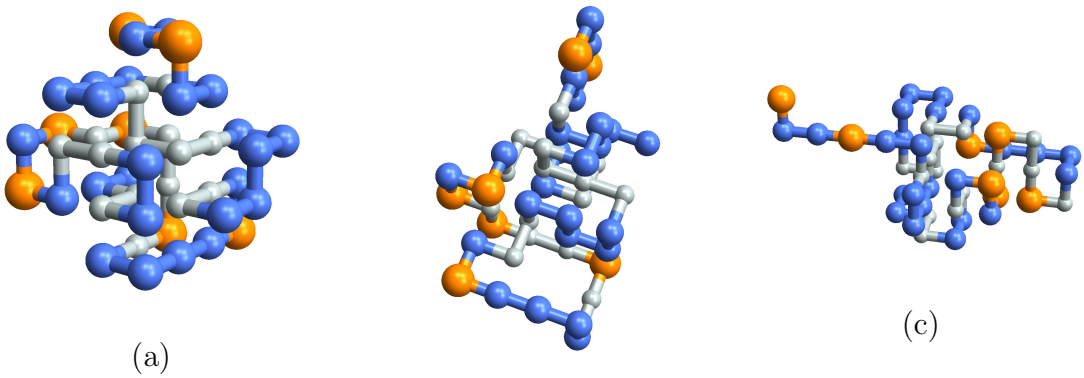
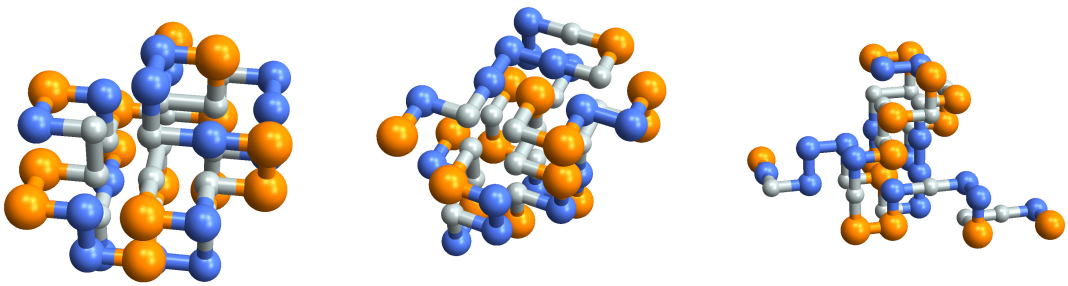


Figure 4.5: Specific Heat (C_V) curves for three H0P lattice homologues *1jmn*, *1orl* and *3c8p* along with the curve for lattice Crambin



(a) (b) (c)

Figure 4.6: Structures of H0P lattice Crambin in different phases (a) folded state; (b) globule state; (c) random coil.



(a) (b) (c)

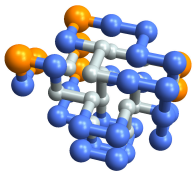
Figure 4.7: Structures of H0P lattice *1wuw* in different phases (a) folded state; (b) globule state; (c) random coil.

Chapter 5

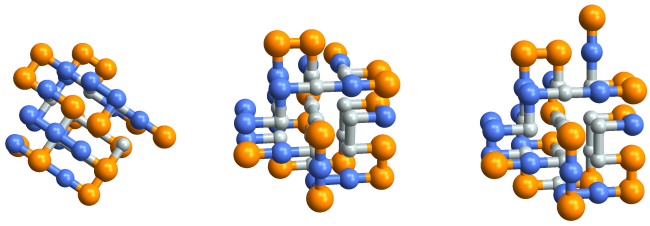
Results: Structural Properties

5.1 Native Structure

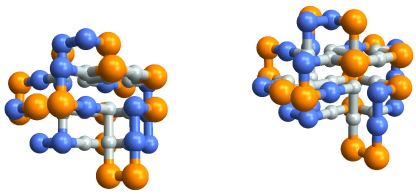
Table 4.1 lists the degeneracy of native states for all lattice homologues tested in this research. The low native state degeneracy in the semi-flexible H0P model permits investigation into individual native structures. Fig. 5.1 shows the native structures for all six lattice homologues. Generally, all these native structures have solid hydrophobic cores and polar shells as expected for common globular proteins. Most lattice homologues do not have same native structures, though their native structures in nature are similar (Fig. 2.3). The only example is that lattice homologues *1orl* and *3c8p* have same native structures in this model.



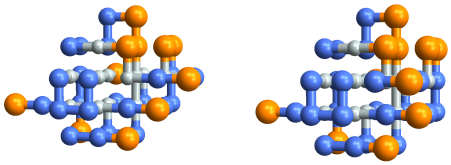
(a) Crambin



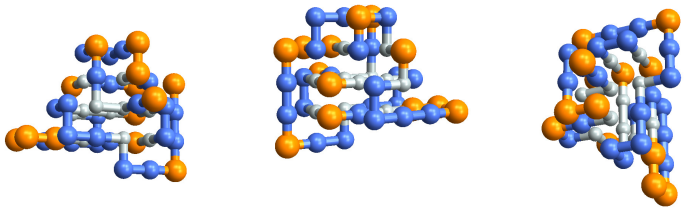
(b) *1bhp*



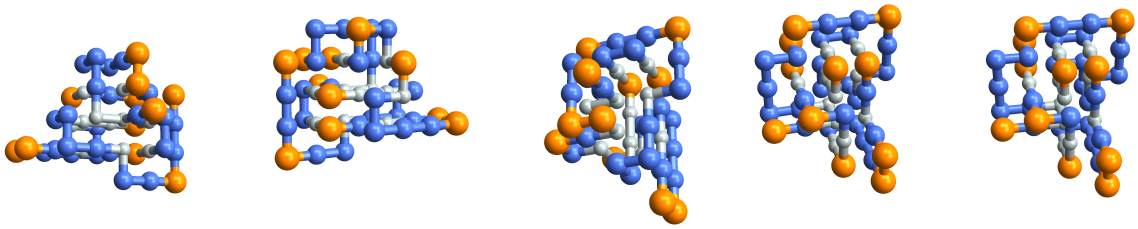
(c) *1wuw*



(d) *1jmn*



(e) *1orl*



(f) *3c8p*

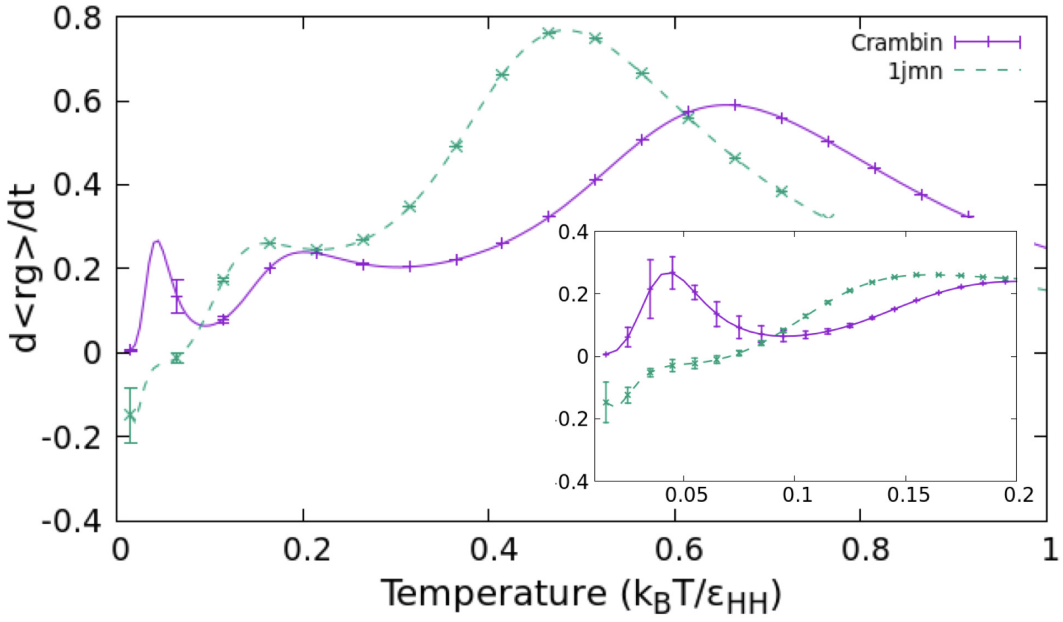
Figure 5.1: Native structures for H0P lattice Crambin and its lattice homologues

5.2 Radius of Gyration

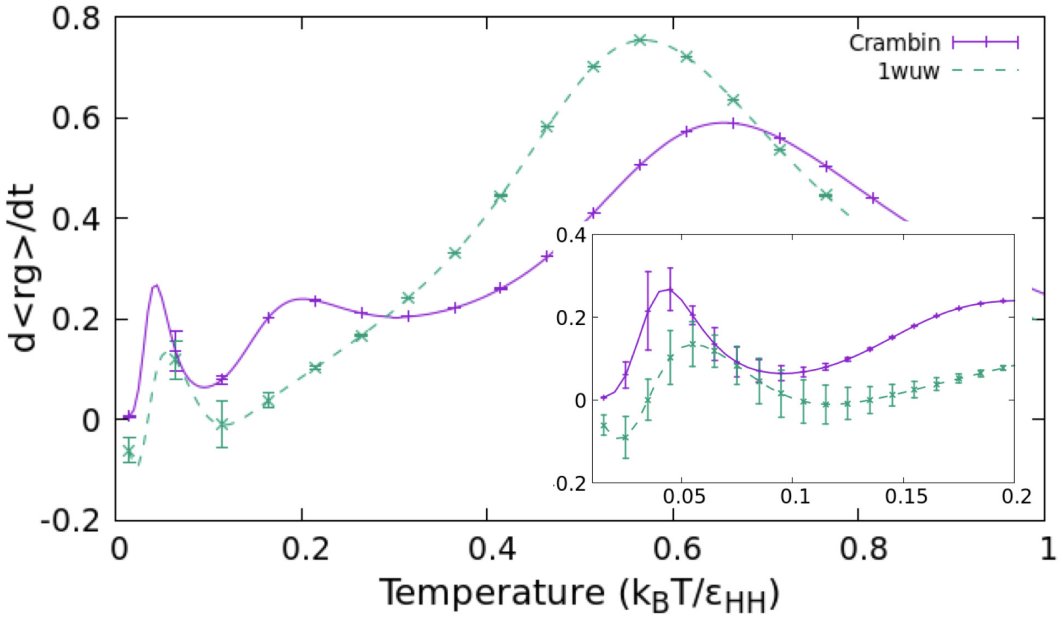
Radius of gyration is a conformational state that shows how compact a polymer folds. During the REMUCA simulation, the radius of gyration is calculated with Eq. 3.2. Two curves of the derivatives of average radii of gyration are shown in Fig. 5.2 as examples, while the results for the remaining homologues can be found in Appendix **A**. Opposite to the thermodynamic behaviors, as we can see in Fig. 5.2(a) and (b), the derivatives of radii of gyration for these lattice homologues have very little in common. Their differences happens not only in the low temperature region as magnified in the graphs, but also in the high temperature region. Although the curves for lattice Crambin and *1jmn* show both folding transitions and coil-globule transitions, for lattice *1wuw* in Fig. 5.2(b), there is only one peak at high temperature, which is very different from other lattice homologues. Therefore, it is hard to compare structural properties in a statistical way for these homologues in the lattice model.

5.3 Contact Map

For most structures of different lattice proteins, the similarity among them is not always straightforward to summarize via merely 3D visualization. Therefore, a more concise approach, the contact map, is applied here to compare different structures. Furthermore, although the introduction of stiffness into the model significantly reduces degeneracies in low lying states [8], in most cases, it is unlikely to have a unique structure in low-lying energy states. So, an average contact map is used here to trace the common features of degenerated states [14].



(a) *1jmn*



(b) *1wuw*

Figure 5.2: Derivatives of average radii of gyration (r_g) for two H0P lattice homologues, compared with the results for Crambin.
Low temperature regions are magnified in the graphs for clarity. The curves for other lattice homologues can be found in Appendix **A**.

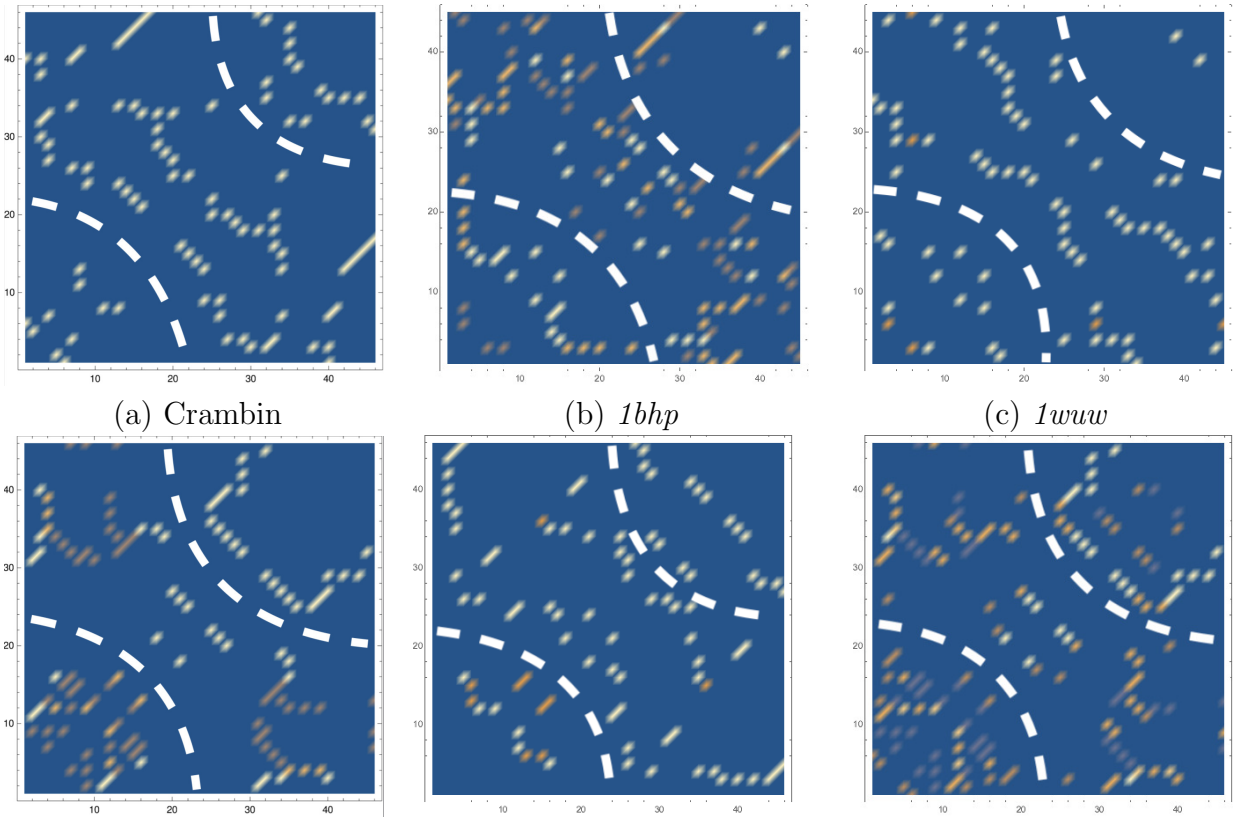
5.3.1 Homology

Due to the constraint of lattice models *per se*, it is usually hard to directly observe features like α-helices or β-sheets, though helix-like structures in the lattice protein model are reported [29]. However, from some contact maps, it is possible to infer that some structural features close to the double-helix structure of these lattice proteins in nature. The average contact maps of native state for all lattice homologues are shown in Fig. 5.3. As indicated by the dashed lines on the contact maps, there is a large potion of contacts concentrated on two ends of the polymers, which are approximately what their natural native structures [32] tell us. The rest contacts along the diagonal are interpreted as interactions between two "helices" in the lattice model.

5.3.2 Evolution

Fig. 4.4 and 4.5 point out some slight configuration change happens at very low temperature for all these lattice homologues. For some lattice homologues, the comparison of contact maps between native state structures and low-lying excited state structures suggests that those low-lying excited states keep the native contacts in the low temperature region. The contact maps for several low-lying excited states of the lattice protein *3c8p* are plotted as an example in Fig. 5.4. By comparing to Fig. 5.3(e), it is obvious that the major part of native contacts remains at low temperature, which implies the "melting" behavior in the low temperature region mostly starts with the rearrangement of bends instead of contacts.

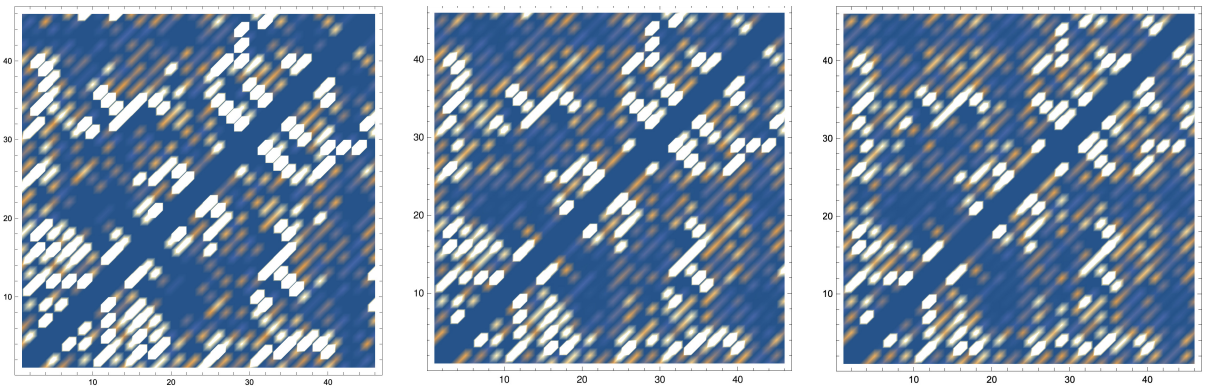
However, some lattice homologues change their structures completely when the temperature slightly increases, as shown in Appendix **B**. Thus, in the semi-flexible H0P model, these lattice homologues do not follow similar folding pathways.



(a) Crambin (b) *1bhp* (c) *1wuw*

(d) *1orl* (e) *1jmn* (f) *3c8p*

Figure 5.3: Average native contact maps for the lattice homologues
(a) native contact map for lattice Crambin; (b) native contact map for lattice *1bhp*; (c) native contact map for lattice *1wuw*; (d) native contact map for lattice *1orl*; (e) native contact map for lattice *1jmn*; (f) native contact map for lattice *3c8p*.
The white dashed lines on the maps indicate contacts inside the "head" and "tail" regions for all lattice homologues. These highlighted regions support that most native states for the lattice homologues have highly compressed structures in two ends.



(a) 2^{nd} state (b) 4^{th} excited state (c) 6^{th} excited state

Figure 5.4: Average contact maps for the low-lying excited states of lattice protein *3c8p* (a) contact map for the 2^{nd} excited state; (b) contact map for the 4^{th} excited state; (c) contact map for the 6^{th} excited state.

Chapter 6

Results: Revisit the Model

All previous results in this work are done with only one value of the stiffness. So, it is worthwhile to test the semi-flexible model with various levels of stiffness. To do this, a 2D REWL simulation with another quantity, the number of "bends" (n_θ), has been done. After getting the 2D density of states ($g(E, n_\theta)$) from simulation, for a given value of ϵ_θ, the 1D density of states can be derived by $g(E_i) = \sum_{\{n_\theta\}} g(E_i, n_\theta)$. Then, the specific heat of the lattice protein system can be calculated at any temperature. Therefore, by changing ϵ_θ, the specific heat curves form a surface. As two examples, the specific heat surfaces for Crambin and $1orl$ are plotted in Fig. 6.1 and Fig. 6.2, respectively, with $(\epsilon_{HH}, \epsilon_{H0})$ set to (2,1). The 2D simulation results for other lattice homologues are shown in Appendix **C**.

Also, a comparison of the low-lying state structures is performed among all ϵ_θ values. The results indicate that, for a range of ϵ_θ, the lattice protein systems keep their low-lying state structures. In the following two figures, different types of curves represent a protein has different ground and low-lying excited state structures with the corresponding value of ϵ_θ.

In both surfaces, there are small shoulders outside the two regular transition signals (coil-globule and folding transition) in the very low temperature region. This corresponds to that

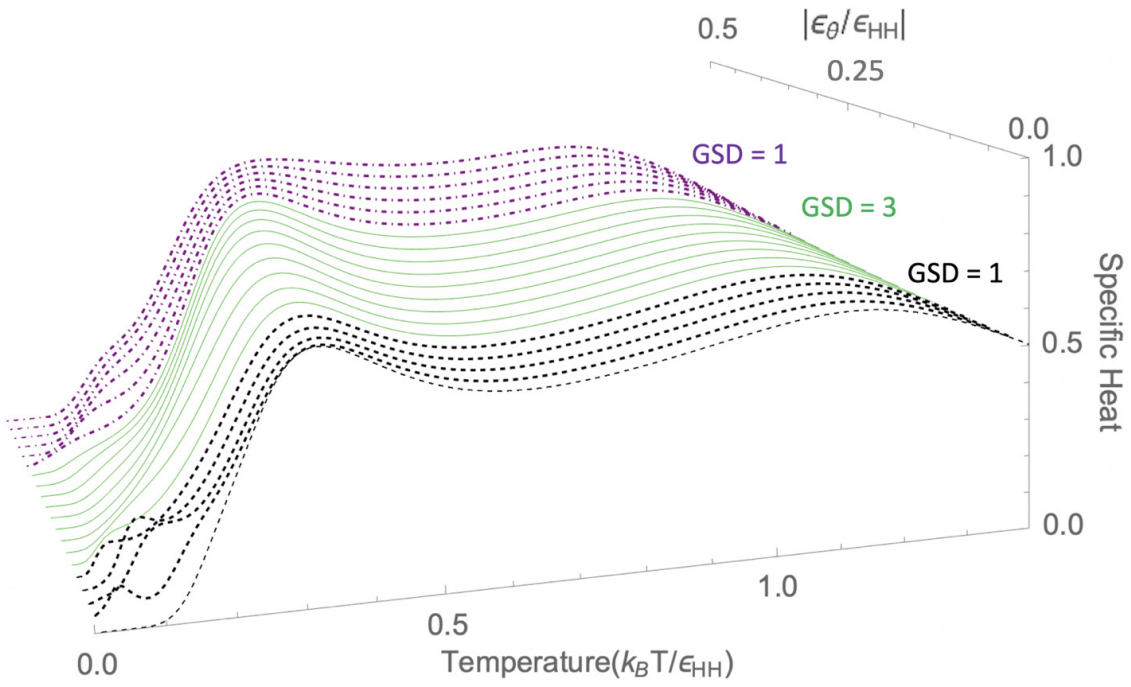


Figure 6.1: Specific heat surface of Crambin obtained from 2D REWL simulation

There are three different sets of ground and low-lying excited states within the whole range of $\epsilon_\theta \in (-\frac{\epsilon_{HH}}{2}, 0)$, marked with different types of curves in the graph. The ground state degeneracy (GSD) of each set is also shown in the graph.

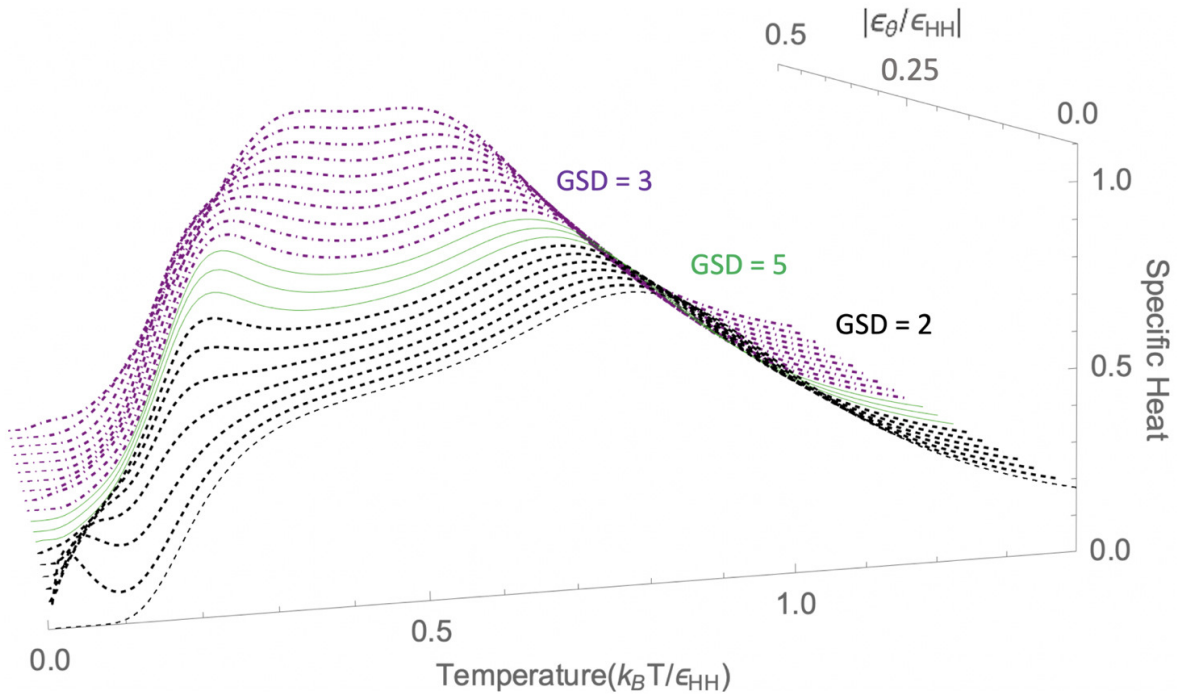


Figure 6.2: Specific heat surface of *1orl* obtained from 2D REWL simulation

There are three different sets of ground and low-lying excited states within the whole range of $\epsilon_\theta \in (-\frac{\epsilon_{HH}}{2}, 0)$, marked with different types of curves in the graph. The GSD of each set is also shown in the graph.

the lattice protein systems are sensitive to small thermal fluctuation at low temperature with the values of interaction chosen.

For 5 of 6 homologues tested in the model, when the value of ϵ_θ is confined to $(0,\epsilon_{H0})$, there are only three different sets of low-lying states. Therefore, these structures of the homologues are stable with slight changes of ϵ_θ, which increases the robustness of the model and the choice of parameters. Meanwhile, for these homologues, with any value of ϵ_θ, the ground state degeneracies are always kept low, which makes it possible for us to have a closer look at the folding behavior of lattice proteins at low temperature.

Chapter 7

Conclusions

To test the homology in the semi-flexible H0P lattice model, Crambin and its homologues are studied in this project as examples. With the help of REWL and REMUCA simulation, thermodynamic and structural quantities are derived at high "resolution" for these lattice homologues.

First of all, for the model *per se*, its performance on these lattice homologues is stable. All through the simulation, the model generates low degeneracy for native and other low-lying states as well as clear two-stage "phase transitions", which are very important for mimicking real proteins in a coarse-grained model. Therefore, the model serves as a powerful tool for understanding protein-like folding behaviors. Moreover, as indicated in Chapter **6**, slight modification of the chain stiffness is unlikely to change the simulation results dramatically, which means the values of interactions can be properly set to small integers (*e.g.* $\epsilon_{HH} = 4$, $\epsilon_{H0} = 2$, $\epsilon_{\theta} = -1$) to avoid wide energy range in the Wang-Landau simulation. This also increases the reliability and efficiency of this model in practice.

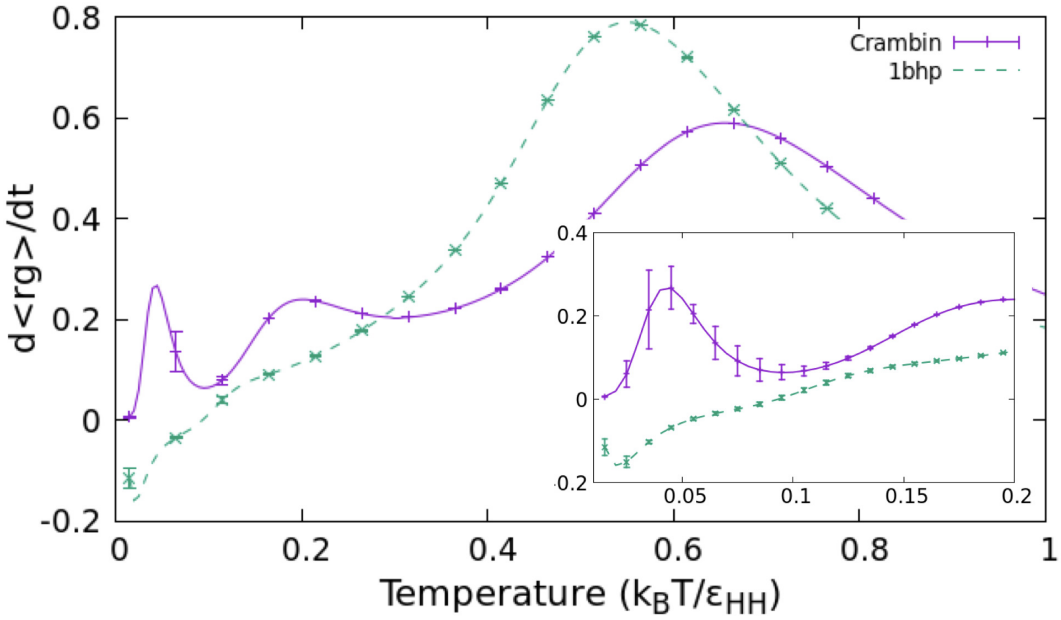
The thermodynamic properties of these lattice homologues in the model are similar. When temperature is very low, the behaviors of these homologues are synchronous with the increase of temperature. The signals in the low temperature region indicates that they

start adjusting their structures at the same temperature. However, the contact maps show that some homologues change their structures slightly during these structural adjustments while others deviate from their native structures thoroughly. This result stops us from implementing the lattice model to study the structural properties on homologues in some way. As the temperature increases, the specific heat curves demonstrate that these lattice homologues undergo folding transitions and coil-globule transitions.

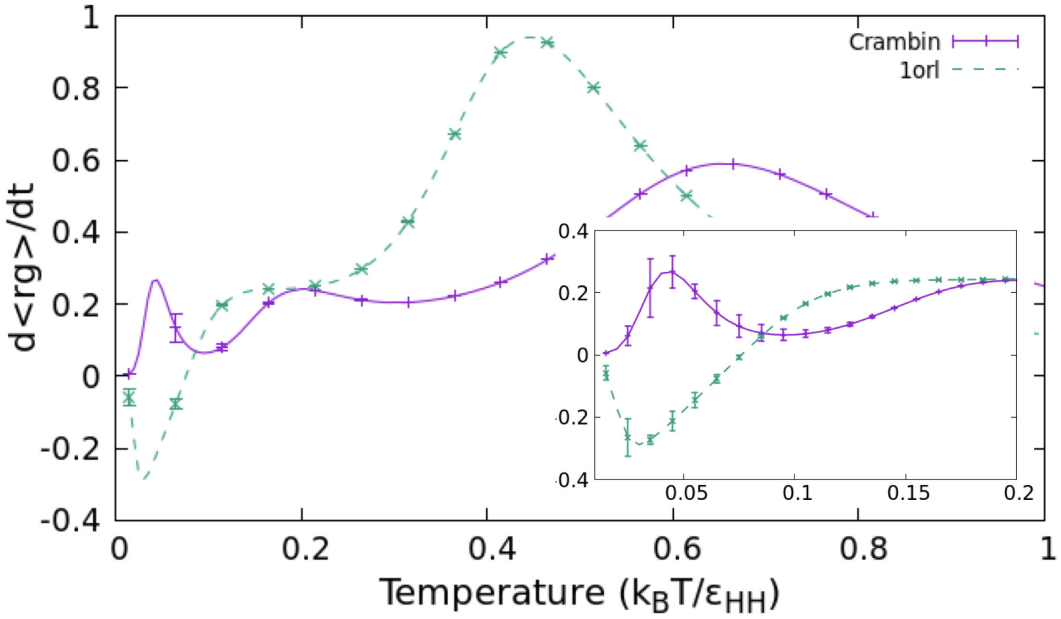
On the other hand, although it is hard to directly spot the similarity between the folded structures of different lattice homologues, some characteristics are still noticeable during the visualization of contact maps. Two groups of contacts concentrate on the "head" and "tail" regions for all these lattice homologues, which are consistent with their "double-helix" structures in nature.

Appendix A

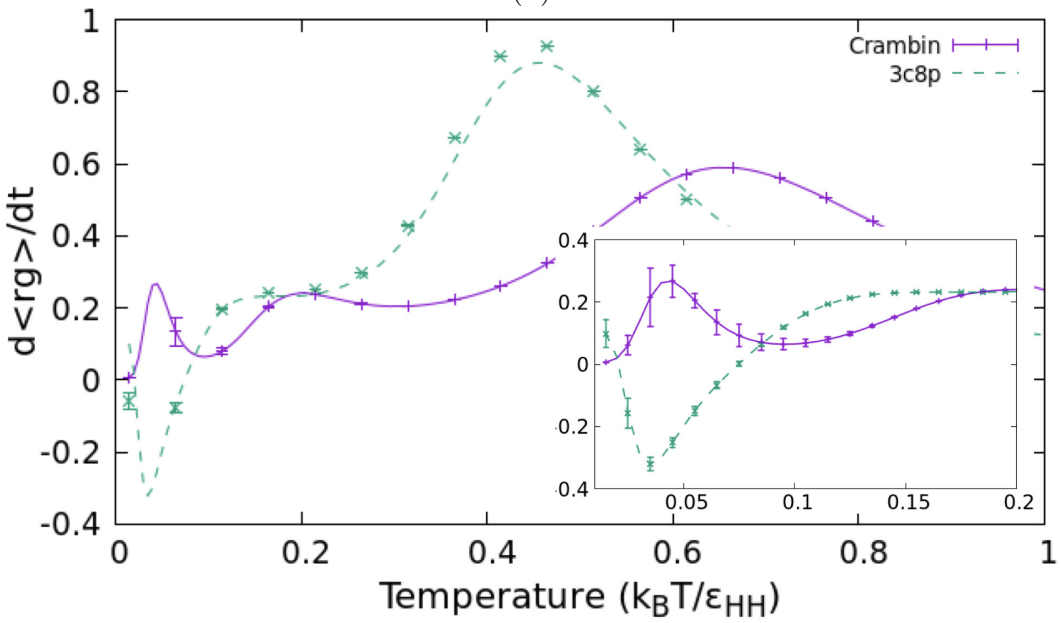
Radius of Gyration



(a) *1bhp*



(b) *1orl*

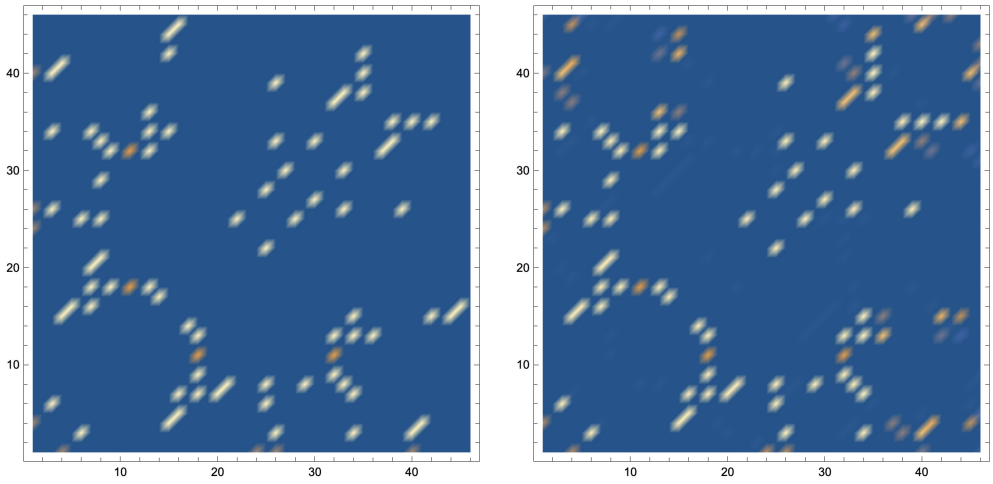


(c) *3c8p*

Figure A.1: Average radius of gyration (r_g) of the remaining lattice homologues, compared with those of Crambin.

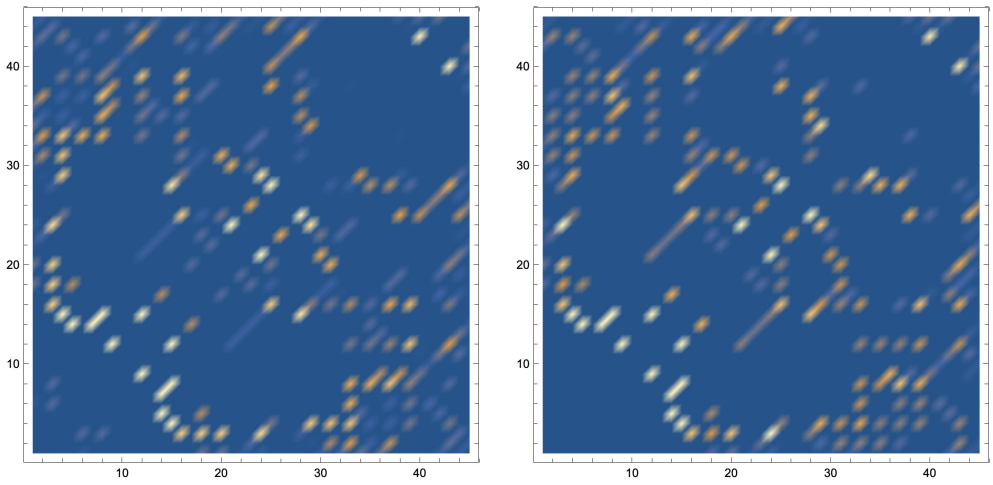
Appendix B

Contact Maps for Low-lying Excited States of Lattice Homologues



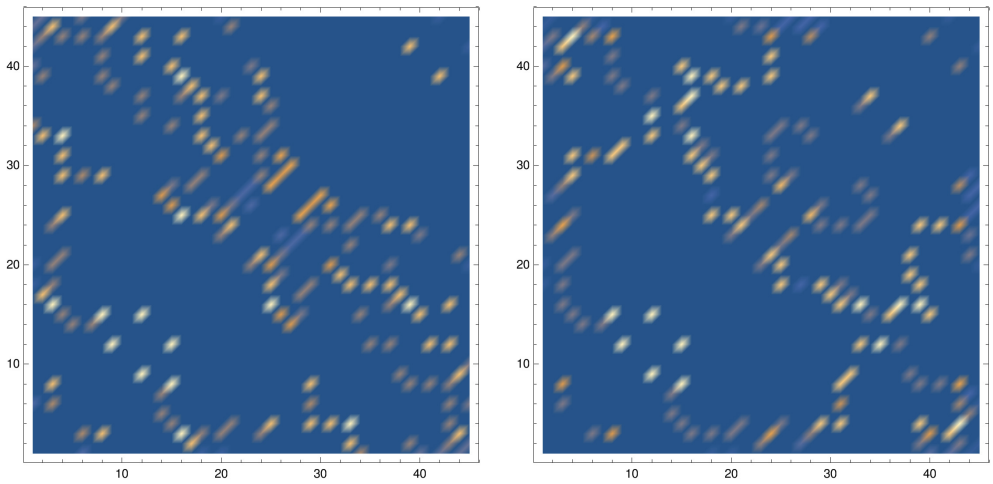
(a) 2^{nd} excited state (b) 4^{th} excited state

Figure B.1: Contact maps for the 2^{nd} and 4^{th} excited state structures of H0P lattice Crambin



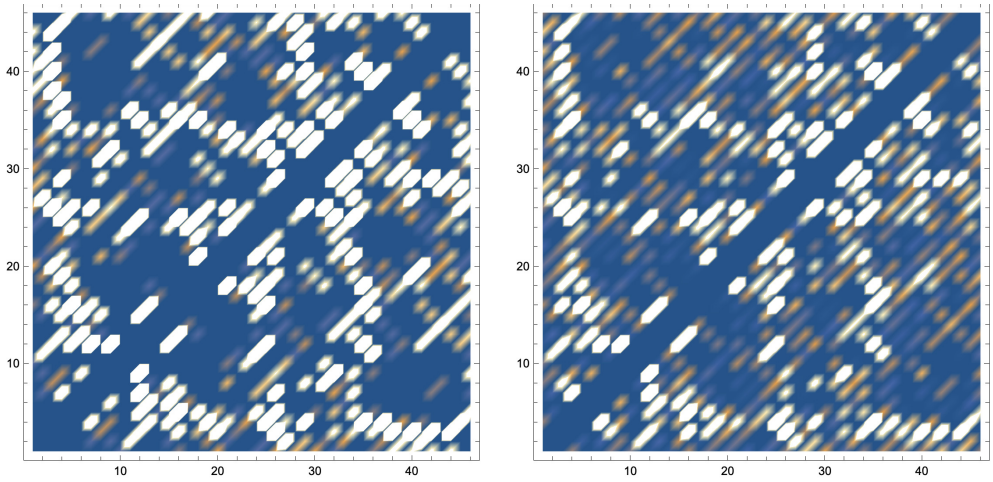
(a) 2^{nd} excitedstate (b) 4^{th} excited state

Figure B.2: Contact maps for the 2^{nd} and 4^{th} excited state structures of H0P lattice *1bhp*



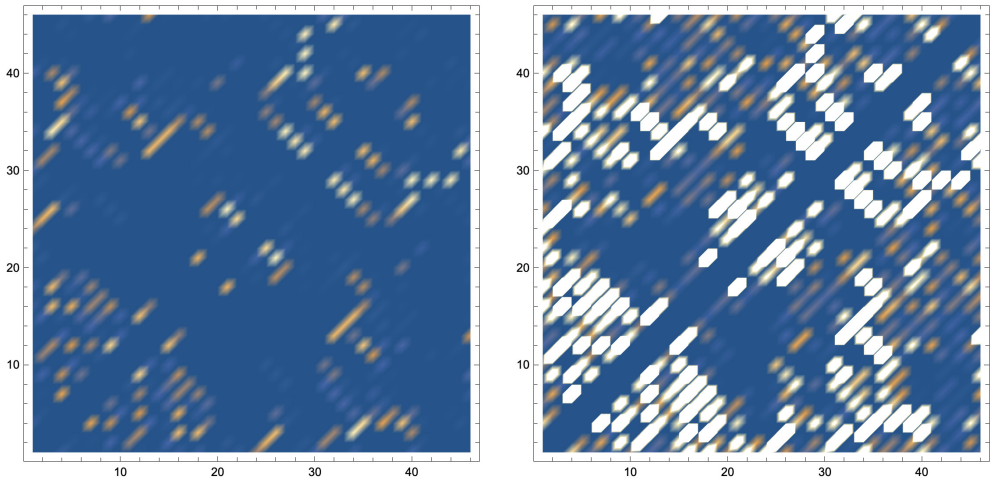
(a) 2^{nd} excited state (b) 4^{th} excited state

Figure B.3: Contact maps for the 2^{nd} and 4^{th} excited state structures of H0P lattice *1wuw*



(a) 2^{nd} excited state (b) 4^{th} excited state

Figure B.4: Contact maps for the 2^{nd} and 4^{th} excited state structures of H0P lattice *1jmn*



(a) 2^{nd} excited state (b) 4^{th} excited state

Figure B.5: Contact maps for the 2^{nd} and 4^{th} excited state structures of H0P lattice *1orl*

Appendix C

2D Density of States

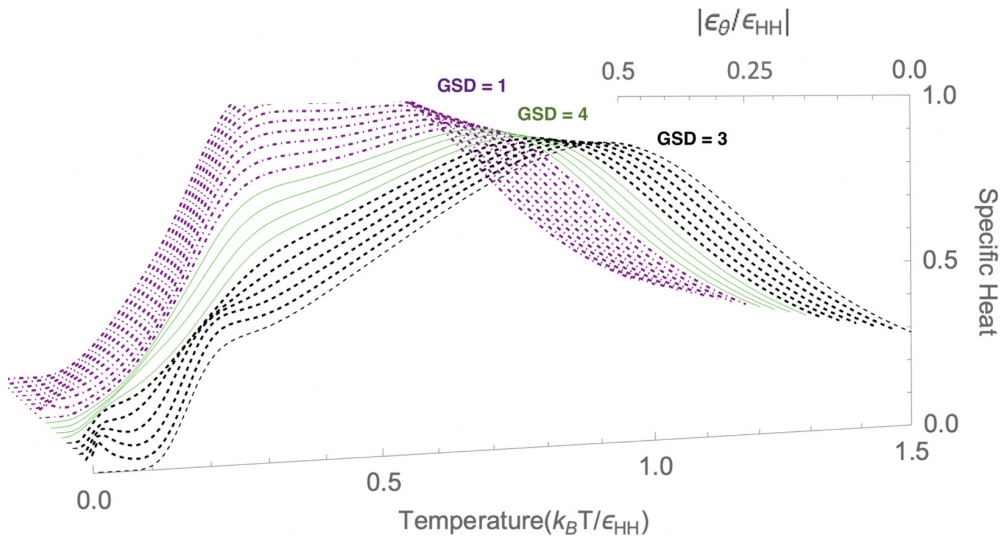


Figure C.1: Specific heat surface of H0P lattice *1bhp* obtained from 2D REWL simulation

There are three different sets of ground and low-lying excited states within the whole range of $\epsilon_\theta \in (0, \epsilon_{H0})$, marked with different types of curves in the graph.

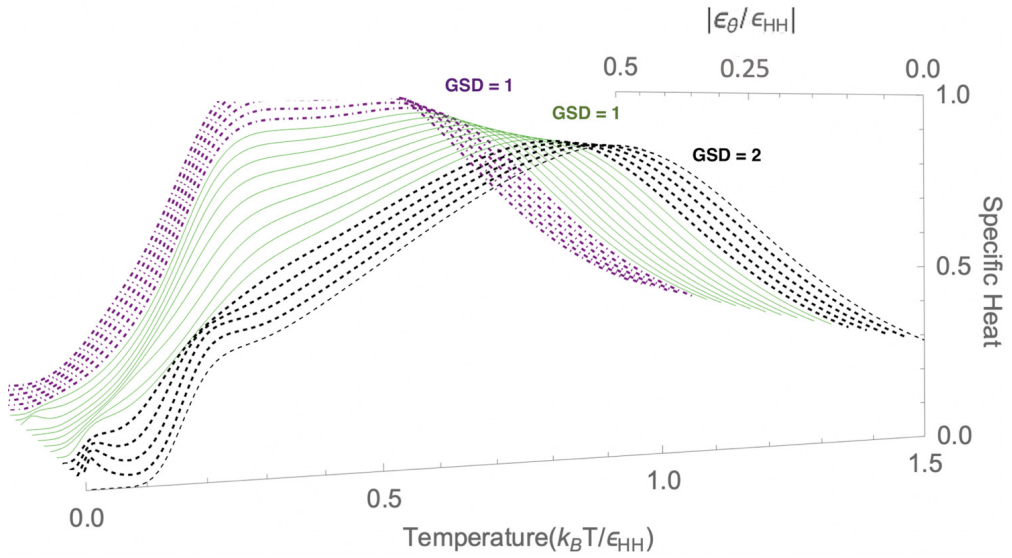


Figure C.2: Specific heat surface of H0P lattice *1wuw* obtained from 2D REWL simulation

There are three different sets of ground and low-lying excited states marked with different types of curves in the graph.

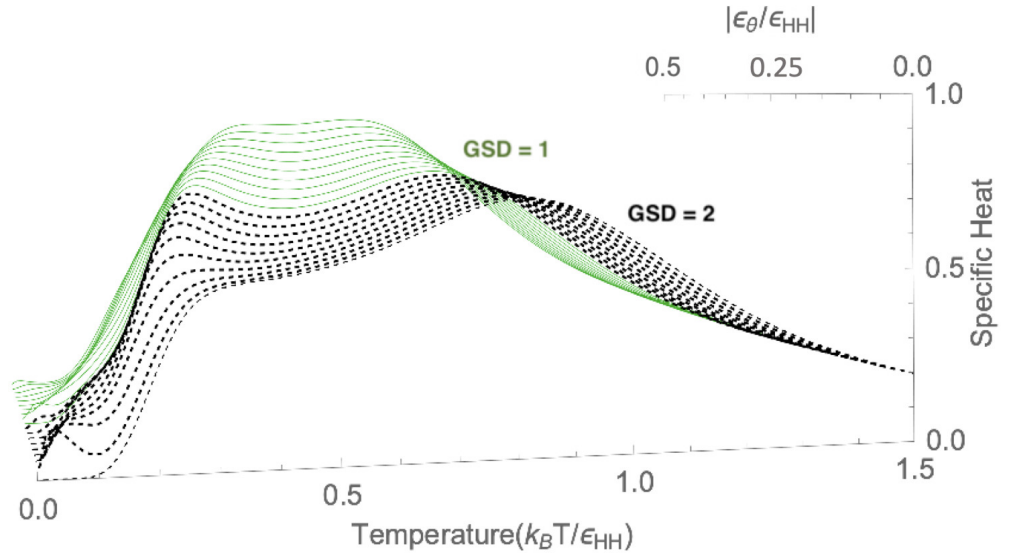


Figure C.3: Specific heat surface of H0P lattice *1jmn* obtained from 2D REWL simulation

There are two different sets of ground and low-lying excited states marked with different types of curves in the graph.

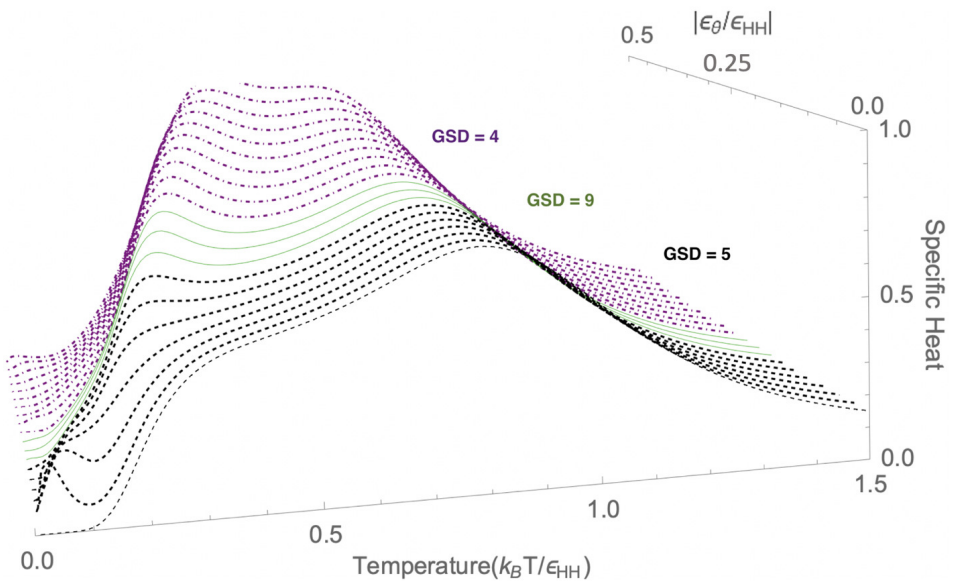


Figure C.4: Specific heat surface of H0P lattice *3c8p* obtained from 2D REWL simulation

There are three different sets of ground and low-lying excited states marked with different types of curves in the graph.

Bibliography

[1] D. P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, 2014).

[2] K. A. Dill and J. L. MacCallum, Science **338**, 1042 (2012).

[3] G. A. Voth, *Coarse-graining of Condensed Phase and Biomolecular Systems* (CRC press, 2008).

[4] D. Fritz, V. A. Harmandaris, K. Kremer, and N. F. van der Vegt, Macromolecules **42**, 7579 (2009).

[5] S. Riniker, J. R. Allison, and W. F. van Gunsteren, Phys. Chem. Chem. Phys. **14**, 12423 (2012).

[6] K. A. Dill, Biochemistry **24**, 1501 (1985).

[7] K. Yue *et al.*, Proc. Natl. Acad. Sci. **92**, 325 (1995).

[8] A. C. Farris, G. Shi, T. Wüst, and D. P. Landau, J. Chem. Phys. **149**, 125101 (2018).

[9] G. Shi, T. Wuest, Y. W. Li, and D. P. Landau, J. Phys. Conf. Ser. **640**, 012017 (2015).

[10] D. Landau, S.-H. Tsai, and M. Exler, Am. J. Phys **72**, 1294 (2004).

[11] Y. W. Li, T. Vogel, T. Wüst, and D. P. Landau, J. Phys. Conf. Ser. **510**, 012012 (2014).

[12] G. Shi, A. C. Farris, T. Wüst, and D. P. Landau, J. Phys. Conf. Ser. **686**, 012001 (2016).

[13] G. Shi, T. Wüst, and P. L. David, J. Phys. Conf. Ser. **905**, 012016 (2017).

[14] G. Shi, T. Wüst, and D. P. Landau, J. Chem. Phys. **149**, 164913 (2018).

[15] A. R. Davidson, Proc. Natl. Acad. Sci. **105**, 2759 (2008).

[16] R. L. Dunbrack Jr, Curr. Opin. Struct. Biol. **16**, 374 (2006).

[17] B. Webb and A. Sali, *Protein Structure Prediction* (Springer, 2014).

[18] N. V. Dokholyan, B. Shakhnovich, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. **99**, 14132 (2002).

[19] A. R. Davidson, K. J. Lumb, and R. T. Sauer, Nat. Struct. Biol. **2**, 856 (1995).

[20] D. M. Taverna and R. A. Goldstein, J. Mol. Biol. **315**, 479 (2002).

[21] F. Morcos *et al.*, Proc. Natl. Acad. Sci. **108**, E1293 (2011).

[22] G. Malgieri *et al.*, Chem. Sci. **9**, 3290 (2018).

[23] P. A. Alexander, Y. He, Y. Chen, J. Orban, and P. N. Bryan, Proc. Natl. Acad. Sci. **104**, 11963 (2007).

[24] D. R. Booth *et al.*, Nature **385**, 787 (1997).

[25] E. Querol, J. A. Perez-Pons, and A. Mozo-Villarias, Protein. Eng. Des. Sel. **9**, 265 (1996).

[26] N. Tokuriki and D. S. Tawfik, Curr. Opin. Struct. Biol. **19**, 596 (2009).

[27] D. Shortle, H. S. Chan, and K. A. Dill, Protein Sci. **1**, 201 (1992).

[28] C. Holzgräfe, A. Irbäck, and C. Troein, J. Chem. Phys **135**, 11B611 (2011).

[29] G. Shi, T. Vogel, T. Wüst, Y. W. Li, and D. P. Landau, Phys. Rev. E **90**, 033307 (2014).

[30] J. D. Bloom, C. O. Wilke, F. H. Arnold, and C. Adami, Biophys. J. **86**, 2758 (2004).

[31] S. Govindarajan and R. A. Goldstein, Proc. Natl. Acad. Sci. **95**, 5545 (1998).

[32] A. Schmidt, M. Teeter, E. Weckert, and V. S. Lamzin, Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. **67**, 424 (2011).

[33] C. Levinthal, Mossbauer Spectroscopy in Biological Systems **67**, 22 (1969).

[34] U. Consortium, Nucleic Acids Res. **43**, D204 (2015).

[35] S. Kmiecik *et al.*, Chem. Rev. **116**, 7898 (2016).

[36] C. Clementi, Curr. Opin. Struct. Biol. **18**, 10 (2008).

[37] G. Ping *et al.*, J. Chem. Phys. **118**, 8042 (2003).

[38] V. Castells, S. Yang, and P. R. Van Tassel, Phys. Rev. E **65**, 031912 (2002).

[39] D. E. Shaw *et al.*, Science **330**, 341 (2010).

[40] A. Koliński *et al.*, Acta. Biochim. Pol. **51** (2004).

[41] A. Liwo *et al.*, J. Mol. Model **20**, 2306 (2014).

[42] J. Honeycutt and D. Thirumalai, Biopolymers **32**, 695 (1992).

[43] H. A. Scheraga, M. Khalili, and A. Liwo, Annu. Rev. Phys. Chem. **58**, 57 (2007).

[44] P. Kar and M. Feig, Adv. Protein Chem. Struct. Biol. **96**, 143 (2014).

[45] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1996).

[46] J. Skolnick, A. Godzik, L. Jaroszewski, and A. Kolinski, Protein Sci. **6**, 676 (1997).

[47] I. Bahar, J. R. L, and K. A. Dill, *Protein Action Principles and Modeling* (Garland Science, 2017).

[48] B. Rost, Protein Eng. **12**, 85 (1999).

[49] M. Stefani and C. M. Dobson, J. Mol. Med. **81**, 678 (2003).

[50] S. Dalal, S. Balasubramanian, and L. Regan, Nat. Struct. Biol. **4**, 548 (1997).

[51] M. Jamroz and A. Kolinski, BMC Struct. Biol. **10**, 5 (2010).

[52] J. Yang *et al.*, Nature Methods **12**, 7 (2015).

[53] N. Yanev, M. Traykov, P. Milanov, and B. Yurukov, J. Comput. Biol. **24**, 412 (2017).

[54] B. Berger and T. Leighton, J. Comput. Biol. **5**, 27 (1998).

[55] S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, and M. H. Hecht, Science **262**, 1680 (1993).

[56] K. A. Dill, Protein Sci. **8**, 1166 (1999).

[57] U. Bastolla and P. Grassberger, J. Stat. Phys. **89**, 1061 (1997).

[58] K. A. Dill, K. M. Fiebig, and H. S. Chan, Proc. Natl. Acad. Sci. **90**, 1942 (1993).

[59] J. É. Debreczeni, B. Girmann, A. Zeeck, R. Krätzner, and G. M. Sheldrick, Acta. Crystallogr. D Biol. Crystallogr. **59**, 2125 (2003).

[60] G. Bhanot, Rep. Prog. Phys. **51**, 429 (1988).

[61] N. Lesh, M. Mitzenmacher, and S. Whitesides, *RECOMB'03 Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology* (ACM New York, 2003).

[62] J. Deutsch, J. Chem. Phys. **106**, 8849 (1997).