Effects of Stiffness on Low Energy States in a Lattice Protein Model for Crambin

Alfred C.K. Farris¹, Guangjie Shi¹, Thomas Wüst², and David P. Landau¹

¹ Center for Simulational Physics, The University of Georgia, Athens, GA 30602, USA
² Scientific IT Services, ETH Zürich, 8092 Zürich, Switzerland

E-mail: alfred.farris@uga.edu

Abstract.
Many studies inspired by the HP lattice protein model have helped to confirm the importance of the hydrophobic “driving force” during folding. Unfortunately, the high level of coarse-graining inherent to this model leads to significant limitations; results from proteins studied under the framework of the HP model fail to reproduce many, sometimes significant, details of the folding process, and the obtained ground states are usually highly degenerate. We propose simple modifications to the original HP model, with the goal of reducing degeneracy and gaining insight into how other interaction parameters influence the folding, while retaining the computational simplicity of lattice models. Namely, we introduce a “neutral” monomer (0) to further divide the hydrophobicity scale and an energetic penalty for “bends” in the protein to account for rigidity. Using replica-exchange Wang-Landau (REWL) sampling and suitable Monte Carlo trial moves, we obtain a unique (non-degenerate) ground state for the new lattice mapping of Crambin (a small, 46 amino acid plant protein), and investigate the effects of stiffness on the folding and the low energy structures.

1. Introduction
Protein folding has been studied from many different perspectives for over half a decade, but many unanswered questions remain; we still do not understand exactly how such complex biomolecules fold so efficiently [1]. The statistical mechanical approach to studying protein folding by way of polymer-based modeling has substantially increased our understanding of these systems [2, 3]. Many studies of, or inspired by, the original hydrophobic-polar (HP) model [4, 5] have highlighted the importance of the hydrophobic “driving force”, but we still wish to uncover which other interactions contribute most relevantly to the folding process [3]. Polymers, as well as polymer models of proteins, such as the aforementioned HP model, have been extensively investigated by a variety of clever Monte Carlo methods [6–20], including Wang-Landau sampling [21,22], which has also been used to successfully study many other problems of biological interest: protein sequence mutation [23,24], folding in a confined environment [25,26], and surface adsorption [27–30].

In addition to the hydrophobic effect, previous studies have shown that rigidity plays an important role in the folding process, as well as in the functional responsibilities of biomolecules [31–35]. We have used replica-exchange Wang-Landau sampling and slight modifications to the
original HP model to investigate the effect of protein stiffness on low energy states of a lattice protein model for Crambin.

2. Models and Simulation Methods

2.1. Semi-flexible H0P Lattice Protein Model for Crambin

The classic HP model greatly simplifies the complexity of protein folding by classifying each of the twenty amino acids as either hydrophobic (H) or polar (P), based on their tendency to interact with water [4, 5]. The protein is then placed on a lattice, with the H and P monomers occupying the lattice sites, connected via rigid bonds. The polymer then performs a self-avoiding random walk, effectively incorporating excluded volume effects inherent to real proteins. To account for the effective hydrophobic driving force, the HP model assigns each non-bonded, nearest neighbor HH contact an energy, \( \epsilon_{HH} \), that lowers the total energy of the system; the configuration(s) with the lowest energy is(are) the ground state(s). One problem inherent to this model is that the ground state energies are degenerate, which stands in contrast to the unique native state found in real proteins. Additionally, this model doesn’t account for polymer stiffness, an effect that should not be overlooked [31–34]. Here we propose a few simple modifications to the original HP model, with the goal of reducing degeneracy and gaining insight into how other interaction parameters influence the folding, while retaining the computational simplicity of lattice models. We first introduce a “neutral” monomer (0) to further divide the hydrophobicity scale [36, 37], and an energetic penalty for “bends” in the protein to account for rigidity [35]; we call this the Semi-flexible H0P model, and a sample configuration, along with interaction parameters is seen in Figure 1. The Hamiltonian for our model is given as follows:

\[
\mathcal{H} = -\epsilon_{HH}n_{HH} - \epsilon_{H0}n_{H0} - \epsilon_{\theta}n_{\theta},
\]

where \( n_{HH} \) is the number of HH contacts, \( \epsilon_{HH} \) is the energy of one such contact, and likewise for the other parameters. In this study, we set \( \epsilon_{HH} = 1 \) and \( \epsilon_{H0} = 0.5 \).

![Figure 1. Sample semi-flexible H0P model configuration. Hydrophobic, neutral, and polar monomers are colored in grey, white, and orange, respectively. The interaction between monomers 2 and 5 contributes energy \( \epsilon_{H0} \), the interaction between monomers 4 and 9 contributes energy \( \epsilon_{HH} \), and the angle formed by monomers 5, 6, and 7 contributes energy \( \epsilon_{\theta} \). In total, there are 2 HH contacts \( (n_{HH}) \), 1 H0 contact \( (n_{H0}) \), and 4 bond-angles \( (n_{\theta}) \).](image)

In this study we investigate Crambin, a slightly hydrophobic protein found in cabbage [38,39]. Due to its relatively small size (46 amino acids), Crambin is one of the more computationally tractable proteins to model. It was previously mapped onto the HP [40] and H0P [37] models with sequence names HP3D46 and H0P3D46, respectively.

2.2. Replica-exchange Wang-Landau Sampling

Traditional Monte Carlo methods, e.g. Metropolis sampling, sample prohibitively slowly at low temperatures and fail for systems with complex free energy landscapes. Wang-Landau sampling [41,42] is an iterative Monte Carlo method whose ultimate goal is obtaining a system’s
density of states (DOS), \( g(E) \), by performing a random walk in energy space. During the simulation, a trial move to change the system from state \( A \) to \( B \) is proposed, and accepted with probability

\[
P(A \rightarrow B) = \min \left\{ 1, \frac{g(E_A)}{g(E_B)} \right\},
\]

where \( g(E_A) \) and \( g(E_B) \) are the current estimates of the DOS for states \( A \) and \( B \), respectively. After a trial move where state \( n \) is accepted, \( g(E_n) \) is updated via \( g(E_n) \rightarrow f \cdot g(E_n) \), where \( f \) is a modification factor. A histogram \( H(E) \) of visited system states is also kept, and updated via \( H(E_n) \rightarrow H(E_n) + 1 \) after state \( n \) is accepted. The histogram is checked at predefined intervals, and reset to zero when all system states have been visited roughly equally: in this study, we enforce an “80% flatness criterion”, meaning all histogram entries must be at least 80% of the average histogram value. The modification factor is then updated via \( f \rightarrow \sqrt{f} \), and this process is repeated until \( f < f_{\min} \), some predefined minimum value.

Replica-Exchange Wang-Landau sampling (REWL) is an extension that combines the power of traditional Wang-Landau sampling with replica-exchange Monte Carlo [43,44]. In REWL, the energy range of a system is split into multiple, overlapping windows, each of which is sampled by one or more independent walkers. A replica-exchange between walkers in neighboring windows \( i \) and \( j \) is proposed periodically, and accepted with probability

\[
P_{\text{acc}} = \min \left\{ 1, \frac{g_i(E_j)}{g_j(E_i)} \right\},
\]

where \( g_i(E_j) \) is walker \( i \)'s current estimate for the density of states of the configuration proposed by walker \( j \), and likewise for the other terms. This framework allows each replica to travel between windows with different energy ranges, which further helps to reduce trapping in free energy minima. REWL produces significant speed-up in simulations of complex systems such as lattice protein adsorption and the Potts model [43,44].

In this study, \( f \) was initially set to \( e^1 \), and \( f_{\min} = \exp(10^{-8}) \). Our trial moves were chosen randomly using the Mersenne Twister random number generator, such that the percentage of each move type were as follows: 75% pull moves [45], 23% bond-rebridging moves [46], and 2% pivot moves [47]. This combination of Monte Carlo moves has proven very efficient in studying lattice proteins together with Wang-Landau sampling [21,22].

2.3. Thermal Properties and Protein Structures

With \( g(E) \) obtained from REWL sampling, the partition function \( (Z) \) can be computed at any temperature \( (T) \) simply as

\[
Z(T) = \sum_E g(E) e^{-E/k_BT},
\]

where \( E \) is the system energy, and \( k_B \) is the Boltzmann constant. We then calculate the statistical mechanical averages of any quantity \( Q \) as follows:

\[
\langle Q \rangle = \frac{\sum_E Q \cdot g(E) e^{-E/k_BT}}{Z}.
\]

The heat capacity \( (C_V) \) can be calculated from fluctuations in the energy via

\[
C_V(T) = \frac{\langle E^2 \rangle - \langle E \rangle^2}{k_BT^2}.
\]
Due to the high degeneracy of most energy levels, as well as the many symmetries associated with the simple cubic lattice, one must be very careful when comparing “different” structures with the same energy. A highly efficient method of searching for, and keeping track of different structures was previously introduced by Shi et. al. [24]. The idea behind this method is to sample configuration space for structures corresponding to a predefined target energy using (replica-exchange) multicanonical sampling (REMUCA) [48, 49], with $g(E)$ from the preceding REWL run as the sampling weight, and translate each structure with this target energy into a sequence of directions (SoD) that accounts for lattice symmetries. During the sampling, each unique SoD is stored in a tree-like data structure, to which newly encountered sequences can be quickly compared. If a given structure is found to be different from all others at that energy, it is added to the tree. This process stops once a certain predefined number of states are found, or if no new structures are found in a certain number of Monte Carlo sweeps. The minimum number of REMUCA sweeps used to calculate structural quantities or obtain ground states for a simulation in this study was $10^9$. Statistical error bars for all quantities in this study were obtained from a minimum of five independent runs.

3. Results

3.1. Specific Heat

![Figure 2](image-url)

Figure 2. Specific heat curves for the HP, H0P, and semi-flexible H0P model for Crambin, with characteristic structures relative to peak locations. Values in the legend correspond to $(\varepsilon_{HH}, \varepsilon_{H0}, \varepsilon_\theta)$. Error bars smaller than the points are not shown.

The specific heat offers insight into a system’s response to thermal excitations, with peaks indicating substantial “transitions”. In smaller proteins, the folding process is thought to occur in two steps [50]: Firstly a coil-globule transition, in which the protein folds from a random coil
to a globule. Secondly, the protein’s structure becomes even more compact, ultimately resulting in a relaxation to the native state. Figure 2 shows the specific heat curve for the HP, H0P, and semi-flexible H0P ($\epsilon_\theta = -0.05$) variants of Crambin, along with characteristic H0P structures relative to the peak locations.

While the original HP model only begins to hint at a second transition (in the form of a broad shoulder), the H0P and semi-flexible H0P models clearly show at least two transitions, consistent with all-atom simulations of Crambin [51]. More interestingly, however, is the emergence of a third structural transition signal, in the form of a small shoulder at very low temperature when we account for bond-angle stiffness. While this shoulder is quite small, it is clearly larger than statistical error bars.

3.2. Ground State Degeneracies and Low Energy Structures

By introducing the “neutral” monomer type, the H0P model reduces the ground state degeneracy to 121 states from $\sim 5.9 \times 10^5$ states in the HP mapping [37]. Adding a small, but opposing energetic contribution ($\epsilon_\theta = -0.05$) for bond-angles further reduces to only a single (unique) native state, consistent with theoretical expectations. The ground state structure is shown in Figure 3. The shoulder that appears at low temperature for this case prompts an investigation into the structures near the ground state.

Figure 3. Ground state structure (a) and bond-only representation (b) for the semi-flexible H0P variant of Crambin with $\epsilon_{HH} = 1$, $\epsilon_{H0} = 0.5$, and $\epsilon_\theta = -0.05$.

Figure 4. First excited states (a, c) and bond-only representation (b, d) for the semi-flexible H0P variant of Crambin with $\epsilon_{HH} = 1$, $\epsilon_{H0} = 0.5$, and $\epsilon_\theta = -0.05$. 
Figure 4 shows the first excited states (a,c), as well as corresponding bond-only representations (b,d). From this figure, it is apparent that the first-excited states are very similar to the ground state, with the only difference being an additional “bend” in the tail end of the protein.

Figure 5. Bond-only representation of the Second excited states for the semi-flexible H0P variant of Crambin with $\epsilon_{HH} = 1$, $\epsilon_{H0} = 0.5$, and $\epsilon_{0} = -0.05$.

Figure 5 shows bond-only representations of the six second-excited states. While these structures are very similar to each other, the internal arrangement of amino acids differs drastically from the lower energy states. A more quantitative study is underway, but we believe that this internal rearrangement corresponds exactly to the low temperature shoulder observed in the specific heat curve.

4. Conclusion

We have used REWL sampling, along with slight modifications to the original HP lattice protein model to investigate the effects of stiffness on folding. By studying the case where $\epsilon_{0} = -0.05$, we reduce the ground state degeneracy by five orders of magnitude from the original HP model, and obtain a non-degenerate ground state. Moreover, we uncover a previously unobserved structural transition at very low temperature. Further investigation of the low energy protein structures reveals a huge change in the internal structure between the second and third excited states. A more detailed study is in preparation, in which we aim to uncover the effects of stiffness on the structural properties during folding, as well as to quantitatively describe the internal rearrangement in structure observed between the second and third excited states in our semi-flexible model.

5. Acknowledgements

This study was supported in part by resources and technical expertise from the Georgia Advanced Computing Resource Center, a partnership between the University of Georgia’s Office
of the Vice President for Research and Office of the Vice President for Information Technology. We would also like to thank the Texas Advanced Computing Resource Center under the NSF XSEDE program, grant no. PHY130014, for computing resources.

References

1. Dill K A and MacCallum J L 2012 Science 338 1042–6
2. Kolinski A and Skolnick J 2004 Polymer 45 511–524
3. Dill K A 1999 Protein Sci. 8 1166–80
4. Dill K A 1985 Biochemistry 24 1501–9
5. Lau K F and Dill K A 1989 Macromolecules 22 3896–3997
6. O’Toole E M and Panagiotopoulos A Z 1992 J. Chem. Phys. 97 8644
7. Beutler T C and Dill K A 1996 Protein Sci. 5 2037–2043
8. Grassberger P 1997 Phys. Rev. E 56 3682–3693
9. Frauenkron H, Bastolla U, Gerstner E, Grassberger P and Nadler W 1998 Phys. Rev. Lett. 80 3149–3152
10. Bachmann M and Janke W 2003 Phys. Rev. Lett. 91 208105
11. Hsu H P, Mehra V, Nadler W and Grassberger P 2003 J. Chem. Phys. 118 444
12. Prellberg T and Krawczyk J 2004 Phys. Rev. Lett. 92 120602
13. Zhang J, Kou S C and Liu J S 2007 J. Chem. Phys. 126 225101
14. Liang F and Wong W H 2001 J. Chem. Phys. 115 3374
15. Zhang J L and Liu J S 2002 J. Chem. Phys. 117 3492
16. Tang W and Zhou Q 2012 Phys. Rev. E 86 031909
17. Unger R and Moult J 1993 J. Mol. Biol. 231 75
18. König R and Dandekar T 1999 BioSystems 50 17–25
19. Shmygelska A and Hoos H H 2005 BMC Bioinf. 6 30
20. Yue K, Fiebig K M, Thomas P D, Chan H S, Shakhnovich E I and Dill K A 1995 PNAS 92 325–329
21. Wüst T and Landau D P 2009 Phys. Rev. Lett. 102(17) 178101
22. Wüst T and Landau D P 2012 J. Chem. Phys. 137 064903
23. Holzgräbe C, Irbäck A and Troein C 2011 J. Chem. Phys. 135 195101
24. Shi G, Vogel T, Wüst T, Li Y W and Landau D P 2014 Phys. Rev. E 90 033307
25. Ping G, Yuan J M, Vallieres M, Dong H, Sun Z, Wei Y, Li F Y and Lin S H 2003 J. Chem. Phys. 118 8042
26. Pattanasiri B, Li Y W, Landau D P, Wüst T and Triampo W 2012 J. Phys.: Conf. Ser. 402 012048
27. Castells V, Yang S and Van Tassel P 2002 Phys. Rev. E 65 031912
28. Swetnam A and Allen M P 2012 Phys. Rev. E 85 062901
29. Radhakrishna M, Sharma S and Kumar S K 2012 J. Chem. Phys. 136 114114
30. Li Y W, Wüst T and Landau D P 2013 Phys. Rev. E 87 012706
31. Thomas P D and Dill K A 1993 Protein Science 2 2050–2065
32. Bastolla U and Grassberger P 1997 J. of Stat. Phys 89 1061
33. Krawczyk J, Owczarek A and Prellberg T 2009 Physica A 388 104–112
34. Williams M J and Bachmann M 2015 Phys. Rev. Lett. 115(4) 048301
35. Shi G, Farris A C K, Wüst T and Landau D P 2016 J. Phys.: Conf. Ser. 866 012001
36. Kyte J and Doolittle R F 1982 J. Mol. Biol. 157 105–132
37. Shi G, Wüst T, Li Y W and Landau D P 2015 J. Phys.: Conf. Ser. 640 012017
38. Vanetten C, Nielsen H and Peters J 1965 Phytochemistry 4 467 – 473 ISSN 0031-9422
39. Hendrickson W A and Teeter M M 1981 Nature 290 107 – 113 ISSN 0031-9422
40. Lattman E E, Fiebig K M and Dill K A 1994 Biochemistry 33 6158–6166
41. Wang F and Landau D P 2001 Phys. Rev. Lett. 86 2050–2053
42. Wang F and Landau D P 2001 Phys. Rev. E 64 1–16
43. Vogel T, Li Y W, Wüst T and Landau D P 2013 Phys. Rev. Lett. 110 210603
44. Vogel T, Li Y W, Wüst T and Landau D P 2014 Phys. Rev. E 90 023302
45. Lesh N, Mitzenmacher M and Whitesides S 2003 RECOMB 188–195
46. Deutsch J M 1997 J. Chem. Phys. 106 8849 ISSN 00219606
47. Sokal A D 1995 Monte Carlo and Molecular Dynamics Simulations in Polymer Science (Oxford University Press) edited by K. Binder
48. Berg B A and Neuhaus T 1991 Physics Letters B 267 249 – 253 ISSN 0370-2693
49. Berg B A and Neuhaus T 1992 Phys. Rev. Lett. 68(1) 9–12
50. Jackson S E 1998 Folding and Design 3 R81 – R91 ISSN 1359-0278
51. Shimada J, Kussell E L and Shakhnovich E I 2001 J. of Mol. Biol. 308 79 – 95