A First Look at Lattice Effects in Coarse-Grained Protein Models via Wang-Landau Simulations

Alfred C.K. Farris¹, Daniel T. Seaton², and David P. Landau¹

¹ Center for Simulational Physics, Department of Physics and Astronomy, The University of Georgia, Athens, GA 30602, USA
² Office for the Vice Provost for Advances in Learning, Harvard University, Cambridge, MA 02138, USA
E-mail: alfred.farris@uga.edu

Abstract.
In order to study the effects of lattice constraints on coarse-grained protein models, we apply Wang-Landau sampling to the continuum analogue of the hydrophobic-polar (HP) lattice protein model. The continuum version is inspired by the AB polymer model but incorporates potentials chosen specifically to mimic those of the lattice case. Because of their relative simplicity, both the lattice and continuum models offer significant computational advantage over all-atom simulations, but the impact of the additional lattice constraint on generic folding behavior is unknown. In this preliminary study, we compare and contrast thermodynamics during the folding process of the continuum model to the original HP lattice protein model for sequences mapped from Crambin, a 46 amino acid plant protein. We find that the folding process for both of these coarse-grained models is quite similar, with major structural transitions occurring at almost exactly the same temperatures.

1. Introduction
Protein folding has been studied for over 50 years, but we still cannot explain how such complex biomolecules fold so efficiently [1]. Coarse-grained models analyzed using statistical mechanics have greatly increased our understanding of protein folding [2], but it is not clear how much detail to include [3, 4]. Thus, one persistent issue is deciding where the ideal balance between model simplicity and physical realism lies. Among coarse-grained models, one substantial difference is that of a lattice vs. continuum representation of the spatial degrees of freedom. The simplification of lattice models allows for efficient computation and more precise interactions, but clearly cannot represent all spatial configurations. Continuum models have the perceived advantage of full spatial resolution, but suffer from the complications associated with empirical potentials and higher computational cost.

The simplest protein models incorporate a reduction of the 20 amino acids into a limited number of types, such as in the hydrophobic-polar (HP) lattice model [5,6], and the continuum AB model [7]. Both of these models use monomers to represent only two distinct amino acids types (H,P or A,B), allowing for a reduction in the different kinds of interactions, but a seemingly substantial difference remains with regard to spatial representation (lattice vs. continuum). It has been difficult to quantify the effects associated with lattice constraints on such models because the potentials used in the HP and AB models are often different enough
that discrepancies in the results can be attributed to the potentials themselves.

In order to efficiently sample the complex free-energy landscape associated with coarse-grained protein models, here we use Wang-Landau sampling [8, 9] to offer a first look into the effects of lattice constraints on coarse-grained models by comparing and contrasting the original HP lattice model to an off-lattice analogue, inspired by the AB model, but for which the potentials are chosen to mimic the lattice case. We compare these models for sequences mapped from Crambin – a small (46 amino acid) plant protein [10, 11]. One note regarding the HP model is that it is an NP-complete optimization problem and has been used as an algorithmic test-bed over a wide range of disciplines; see [12] for an overview of these studies.

2. Models and Simulation Methods

2.1. HP Lattice Model

The HP model simplifies the complexity of protein folding by classifying each of the twenty amino acids as either hydrophobic (H) or polar (P), based on their desire to interact with water [5, 6]. The protein is then modeled as a polymer performing a self-avoiding random walk with H and P monomers placed at sites of a simple cubic lattice, automatically accounting for excluded volume effects found in real proteins. To model the hydrophobic “driving force”, each non-bonded, nearest neighbor HH contact contributes energy $\epsilon_{HH}$, resulting in the following Hamiltonian:

$$H = -\epsilon_{HH}n_{HH},$$

(1)

where $n_{HH}$ is the number of non-bonded HH contacts. Thus, each such contact lowers the internal energy of the system, and leads to the formation of a hydrophobic core in the native state, as in real proteins. The major advantages of this model are that it allows for efficient sampling of conformation space and avoids the uncertainty associated with empirical potentials. Fig. 1 shows a sample HP model configuration with energy $E = -3$ based on three HH interactions.

2.2. HP Model Continuum Analogue

The AB continuum protein model shares many similarities with the HP model, such as the same reduced amino acids and simple interactions. However, the potentials in the AB model are not constructed to mimic the HP model case, and thus side-by-side comparisons are never made. In this study we introduce a continuum analogue to the HP model, where non-bonded hydrophobic-hydrophobic interactions are modeled via a Lennard-Jones potential [13], bonds are connected via a finitely-extensible-nonlinear-elastic (FENE) potential [14], and all other monomer interactions are purely repulsive – using the repulsive component of the Lennard-

Figure 1. Sample HP lattice model configuration. Hydrophobic and polar monomers are colored in gray and orange, respectively. The interaction between non-bonded hydrophobic monomers, such as monomers 4 and 9, contributes energy $\epsilon_{HH}$. In total, there are 3 HH contacts ($n_{HH}$).
Figure 2. Sample continuum HP model configuration. Hydrophobic and polar monomers are colored in gray and orange, respectively. The interaction between non-bonded hydrophobic monomers, such as monomers 4 and 9, contributes energy $\epsilon_{HH} V_{LJ}$. The interaction between non-bonded, non-hydrophobic monomers, such as monomers 1 and 6, contributes energy $V_{EV}$. The interaction between bonded monomers, such as monomers 8 and 9, contributes energy $V_{FENE}$.

Jones potential, to mimic the excluded volume effects inherent to the lattice HP model. Thus, our continuum Hamiltonian is given as follows:

$$H = \epsilon_{HH} \sum_{\langle iH,jH \rangle} V^LJ(r_{ij}) + \sum_{\langle \text{bonded} \rangle} V^{FENE}(r_{\text{bond}}) + \sum_{\langle \text{other} \rangle} V^{EV}(r_{\text{other}}),$$

(2)

where

$$V^LJ(r_{ij}) = \left( \frac{r_0}{r_{ij}} \right)^{12} - 2 \left( \frac{r_0}{r_{ij}} \right)^6,$$

(3)

$$V^{FENE}(r_{\text{bond}}) = -0.5k(r_{\text{max}} - r_0)^2 \ln \left( 1 - \left( \frac{r_{\text{bond}} - r_0}{r_{\text{max}} - r_0} \right)^2 \right),$$

(4)

and

$$V^{EV}(r_{\text{other}}) = \left( \frac{r_0}{r_{\text{other}}} \right)^{12},$$

(5)

where the values $\epsilon_{HH}$ and $r_0$ set the energy and length scales of our system, respectively, and in this study are both set to 1 using dimensionless units, $r_{ij}$ is the distance between non-bonded monomers $i$ and $j$, $r_{\text{bond}}$ is the distance between bonded monomers, and $r_{\text{other}}$ is the distance between non-bonded non-hydrophobic pairs. We choose for the stiffness $k = 120 \frac{\epsilon}{r_0^2}$ (to achieve similar stiffness as in [15]), and for the the maximum displacement from equilibrium $r_{\text{max}} = 1.2r_0$. Moreover, we use a truncated, shifted Lennard-Jones potential such that $V^LJ(r \geq r_c) = 0$, where the cutoff distance $r_c = 3.0r_0$. One important note is that the reference energy value, $\epsilon_{HH}$, is set to 1 in both the lattice and continuum case. Fig. 2 shows a sample continuum HP model configuration and Fig. 3 shows a plot of the continuum potentials used in this study. In contrast to a study using a similar model investigating homopolymer systems, where bonds were connected via a combination of FENE and Lennard-Jones potentials [15], we have found the quantitative differences to be quite small when using only the FENE potential between bonds and thus follow the latter approach here for simplicity.

In this study we investigate Crambin, a slightly hydrophobic protein found in cabbage [10,16]. Due to its relatively small size ($N = 46$ amino acids), Crambin is one of the more computationally tractable proteins to model. Here we use the same HP model mapping as in [17].

2.3. Wang-Landau Sampling

Traditional Monte Carlo methods like Metropolis sampling perform poorly at low temperatures and fail for systems with complex free energy landscapes. Wang-Landau sampling
Figure 3. Continuum HP model potentials. Non-bonded hydrophobic-hydrophobic interactions are given via a Lennard-Jones potential ($V_{LJ}$), bonds are connected via a FENE potential ($V_{FENE}$), and all other monomer interactions are purely repulsive ($V_{other}$ – the repulsive $r^{-12}$ term from the Lennard-Jones potential).

[8,9] is an iterative Monte Carlo method that obtains 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. At this point, a histogram $H(E)$ of visited system states updated via $H(E_n) \rightarrow H(E_n) + 1$ after state $n$ is accepted. The histogram is checked at regular intervals, and reset to zero when all system states have been visited such that all histogram entries are 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. Since the Wang-Landau acceptance probability is independent of temperature, this sampling scheme is much more efficient at sampling systems with complex free energy landscapes.

In this study, $f$ was initially set to $e^1$, and $f_{\min} = \exp(10^{-6})$. In the lattice model, our trial moves were chosen randomly such that the percentage of each move type were as follows: 75% pull moves [18], 23% bond-rebridging moves [19], and 2% pivot moves [20]. This combination of Monte Carlo moves has proven very efficient in studying lattice proteins together with Wang-Landau sampling [12,21]. In the continuum model, we use $N$ diffusion moves for one of each of the following: reptation, bond-rebridging, crankshaft, and random pivot moves.

2.4. Thermal Properties and Protein Structures

With $g(E)$ obtained from Wang-Landau 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 (= 1 in this study). We then calculate the ensemble average 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

$$C_V(T) = \frac{\langle E^2 \rangle - \langle E \rangle^2}{k_BT^2}.$$
3. Results

3.1. Specific Heat

Figure 4. Specific heat curves for the lattice and continuum HP models for Crambin, with characteristic structures shown relative to peak locations. Error bars smaller than the points are not shown.

The specific heat ($C_V/N$) offers insight into the folding process, with peaks indicating structural transitions. In smaller proteins, the collapse process occurs in two steps [22]: Firstly a coil-globule transition, in which the protein folds from a random coil to a globule. Secondly, the “folding transition” consists of the protein becoming even more compact, and ultimately relaxing to the native state. Fig. 4 shows the specific heat curve for both the lattice and continuum HP models of Crambin, along with characteristic structures relative to the peak locations.

Both models exhibit these two structural transitions (consistent with all-atom simulations of Crambin [23]): the lattice HP model displays a peak and broad shoulder, but in the continuum case they both appear as peaks. While the magnitude of these curves differ between models, the temperature scale is common, and thus the “transition” temperatures (and overall results) are strikingly similar. Note that although the specific heat curve for the continuum model appears to go to zero at low temperature, this is a non-physical result of cutting the sampled energy range off at a certain minimum. The true specific heat will level off to a value given by the equipartition theorem at zero temperature. In the lattice case, the drop-off to zero is because
there is a gap between the ground-state and first-excited state energy.

The main difference between the two models is the emergence of a faint transition signal in the form of a small shoulder between the two peaks in the continuum case. We have not found substantial numerical evidence in any order parameter to explain this shoulder (no signal in the radius of gyration, radial distribution function, end-to-end distance, or core density), and thus this shoulder still remains somewhat mysterious. Its origin is currently under investigation and will be included in a further study.

4. Conclusion

We have used Wang-Landau sampling to offer a first look into lattice effects on coarse-grained protein models by examining the original HP model and its continuum analogue for a sequence mapped from Crambin. Analysis of the specific heat curves reveals that the folding process shares many similarities between the two models. The coil-globule transition occurs at almost exactly the same temperature and is marked by a peak in both cases. While the specific heat displays quantitative differences at the folding transition (peak vs. shoulder), the qualitative folding behavior is the same and the temperature at which this transition occurs is similar for both models. This suggests that the lattice constraint does not have a substantial impact on the generic folding behavior in this case. A more detailed study is in progress in which we aim to uncover the quantitative meaning of the small shoulder that appears in the continuum case, as well as investigate lattice effects on semi-flexible protein models.

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.

References

[1] Dill K A and MacCallum J L 2012 Science 338 1042–6
[2] Dill K A and Chan H S 1997 Nat. Struct. Mol. Biol. 4 10–19
[3] Dill K A, Bromberg S, Yue K, Chan H S, Fiebig K M, Yee D P and Thomas P D 1994 Protein Sci. 4 561–602
[4] Kolinski A and Skolnick J 2004 Polymer 45 511–524
[5] Dill K A 1985 Biochemistry 24 1501–9
[6] Lau K F and Dill K A 1989 Macromolecules 22 3986–3997
[7] Stillinger F H, Head-Gordon T and Hirshfeld C L 1993 Phys. Rev. E 48(2) 1469–1477
[8] Wang F and Landau D P 2001 Phys. Rev. Lett. 86 2050–2053
[9] Wang F and Landau D P 2001 Phys. Rev. E 64 056101
[10] Hendrickson W A and Teeter M M 1981 Nature 290 107 – 113 ISSN 0031-9422
[11] Schmidt A, Teeter M, Weckert E and Lamzin V S 2011 Acta Crystallographica Section F 67 424–428
[12] Wüst T and Landau D P 2009 Phys. Rev. Lett. 102(17) 178101
[13] Lennard-Jones J E 1924 Proc. Royal Soc. Lond. A 106 463–477
[14] Kremer K and Grest G S 1990 J. Chem. Phys. 92 5057–5086
[15] Seaton D T, Wüst T and Landau D P 2010 Phys. Rev. E 81(1) 011802
[16] Vanetten C, Nielsen H and Peters J 1965 Phytochemistry 4 467 – 473 ISSN 0031-9422
[17] Lattman E E, Fiebig K M and Dill K A 1994 Biochemistry 33 6158–6166
[18] Lesh N, Mitzenmacher M and Whitesides S 2003 RECOMB 188–195
[19] Deutsch J M 1997 J. Chem. Phys. 106 8849
[20] Sokol A D 1995 Monte Carlo and Molecular Dynamics Simulations in Polymer Science (Oxford University Press) edited by K. Binder
[21] Wüst T and Landau D P 2012 J. Chem. Phys. 137 064903
[22] Jackson S E 1998 Folding and Design 3 R81 – R91 ISSN 1359-0278
[23] Shimada J, Kussell E L and Shakhnovich E I 2001 J. Mol. Biol. 308 79 – 95