Folding in a semi-flexible lattice model for Crambin

Guangjie Shi\textsuperscript{1}, Alfred C.K. Farris\textsuperscript{1}, Thomas Wüst\textsuperscript{2} and David P. Landau\textsuperscript{1}

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

E-mail: sgjerry@physast.uga.edu

Abstract.
Using the Replica-Exchange Wang–Landau sampling method, we investigated and compared three different coarse-grained lattice protein models for the small, hydrophobic protein Crambin. We show that slight extensions of the HP lattice protein model, including the stiffness of bonds can lead to a significant decrease in ground-state degeneracies (up to 5 orders of magnitudes). Moreover, the ground-state structures begin to bear resemblance to native structures observed in real Crambin.

1. Introduction
Generic, coarse-grained protein models have served as great tools to improve our understanding of many important biological problems \cite{1, 2}. Among these models, the hydrophobic-polar (HP) lattice protein model \cite{3, 4} is arguably the simplest, yet it still remains computationally challenging. Finding the lowest energy structures of given sequences in the HP model has been proven to be an NP-complete problem \cite{5}, although different enumeration methods and stochastic algorithms have been developed for effectively approaching this problem. Examples include sequential importance sampling \cite{6}, fragment regrowth Monte Carlo \cite{7}, chain-growth methods \cite{8, 9}, e.g. the pruned-enriched Rosenbluth method (PERM) and its variants \cite{10, 11, 12, 13, 14}, multidomain sampler \cite{15}, genetic algorithms \cite{16, 17}, evolutionary Monte Carlo \cite{18}, ant colony models \cite{19} and Wang–Landau sampling \cite{20, 21}. With these latter methods, several problems of biological interest have been successfully investigated, such as the effects of sequence mutations \cite{22, 23}, protein folding in confined environments \cite{24, 25} and protein surface adsorption \cite{26, 27, 28, 29}.

However, the simplicity of the HP model yields large ground-state degeneracies which stands in contrast to the generally unique native states of natural proteins. To overcome this issue, and also reflect more closely the range of hydrophobicities found among different amino acids a third “neutral” amino acid, noted by “0” has been introduced. The resultant H0P model generally shows less ground-state degeneracies compared to the original HP model \cite{30}. In this paper we shall use Replica Exchange Wang-Landau (REWL) simulations to study the behavior of the protein Crambin mapped onto an H0P model. In addition, we take account of the natural rigidity of real proteins by introducing an energy term that reflects the cost of producing a bond angle between successive bonds. The purpose of this study is to investigate whether the introduction of these additional degrees of freedom may yield a significantly better representation.
of real proteins (here Crambin) while still remaining within the frame of minimalist models with all their advantages (e.g. rapid sampling of conformational space).

2. Models

2.1. H0P lattice protein model

Analogous to the HP model, the H0P model classifies amino acids into three types: hydrophobic (H), “neutral” (0) and polar (P), based on their hydrophobicity levels [31]. The primary structure of a protein is composed of a chain of connected (bonded) beads, each of which represents a single monomer, on a simple cubic lattice. Two non-bonded interactions are considered: $\epsilon_{HH}$ and $\epsilon_{H0}$, where $\epsilon_{HH}$ denotes the interaction strength between non-bonded H-mers on nearest neighbor sites ($\epsilon_{H0}$ has corresponding meaning), and $\epsilon_{HH} > \epsilon_{H0}$. The Hamiltonian is then given by

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

where $n_{HH}$ is the number of non-bonded HH contacts ($n_{H0}$ has corresponding meaning). Other interactions involving “0” monomers could be easily added.

2.2. Semi-flexible H0P model

The idea of considering bond-stiffness energy has been studied previously, e.g., Thomas and Dill explored the relationship between helical propensities and conformations for globular proteins using an HP model with local helix interaction [32]; Bastolla and Grassberger [33] studied a lattice model of semi-flexible homopolymers with nearest neighbor attraction and energetic preference for straight joints between bonded monomers; Krawczyk et al. [34] have extensively investigated semi-flexible hydrogen-bonded and non-hydrogen bonded lattice polymers in both two and three dimensional lattices.

In this work, we include an energetic term $\epsilon_\theta$ for each bent bond joint (i.e. angle) inside the protein structures resulting in the following Hamiltonian:

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

where $n_\theta$ represents the number of angles in a given protein structure. Figure 1 illustrates the various energy terms for a short semi-flexible H0P model.

**Figure 1.** An example of a semi-flexible H0P model. Hydrophobic and “neutral” monomers are colored in gray and white, respectively, while polar monomers are colored in orange. The interaction between monomers 2 and 5 is $\epsilon_{H0}$, and that between monomers 4 and 9 is $\epsilon_{HH}$. The angle constituted by monomers 5, 6 and 7 contributes $\epsilon_\theta$ energy. In this particular 2-dimensional structure, $n_{HH} = 2$, $n_{H0} = 1$ and $n_\theta = 4$.

3. Methods

3.1. Wang–Landau sampling

For systems with complex free energy landscapes, e.g., lattice proteins, traditional Monte Carlo methods such as Metropolis sampling are known to have trouble in sampling the low temperature
region because of trapping in local minima. Wang–Landau (WL) sampling [35, 36] has the ultimate goal of estimating the density of states \( g(E) \) iteratively by performing a random walk in a given energy range \([E_{\text{min}}, E_{\text{max}}]\) and effectively eliminates trapping. Together with suitable Monte Carlo trial moves, WL sampling has proven to be extremely powerful for studying this type of systems [20, 21]. During the simulation, each Monte Carlo trial move that changes the system from energy \( E_A \) to \( E_B \) will be accepted with probability

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

(3)

where \( g'(E) \) is an instantaneous estimator for the density of states. After each trial update \( g'(E_i) \) is updated via \( g'(E_i) \rightarrow f \times g'(E_i) \), where \( i \) refers to the state that results from the trial, \( f \) is a modification factor, and the histogram of visited energies is increased by one. The modification factor is reduced when the histogram is sufficiently “flat” and at the same time, all the entries in the histogram are reset to zero. The procedure is repeated until the modification factor is less than some predefined threshold value \( f_{\text{final}} \). In this work, the initial modification factor is set to \( f_{\text{init}} = e^1 \) and decreased via \( f \rightarrow \sqrt{f} \), and \( f_{\text{final}} = 1 \times 10^{-8} \). Initially, all entries in \( g'(E) \) are assigned to 1 and we apply the “80%” flatness criterion for the histogram, i.e., every entry in the histogram is no less than the 80% of the mean histogram height. As Monte Carlo trial moves, we have used pull and bond-rebridging moves which have been proven to be very efficient together with WL sampling [20, 21].

3.2. Replica-Exchange Wang–Landau sampling
Replica-Exchange Wang–Landau (REWL) [37, 38] sampling, a parallel version of Wang–Landau sampling, fully exploits and combines the power of WL sampling and replica-exchange Monte Carlo. It divides the energy range \([E_{\text{min}}, E_{\text{max}}]\) of the system into multiple, overlapping sub-ranges (windows). Configurations (replicas) in each of the resulting windows are simulated by independent processes (random walkers, running serial sampling, each on a different core), and then replica exchanges between overlapping windows are attempted at fixed time intervals during the simulation. As a result, each replica can travel through the entire energy space of interest. The probability of accepting a replica exchange move between random walker \( i \) and \( j \) from neighboring windows is given by

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

(4)

where \( g_i \) and \( E_i \) denote the estimator for the density of states and the energy of current configuration of walker \( i \), respectively. REWL exhibits very good strong and weak scaling properties for different systems including the Potts model, lattice protein adsorption, or the self-assembly process in amphiphilic solutions [37, 38] and is very effective in speeding up our simulations.

3.3. Ground state sampling
The problem of finding the ground-state structures of lattice protein models is NP-complete [5]. And due to the exponential growth of combinations, exact enumeration methods become unfeasible for protein sequences longer than 30 monomers. However, in previous work we proposed a ground state sampling method [23] that relies on Wang–Landau sampling to obtain reliable estimates of the ground-state degeneracies. The idea behind our method is to perform multicanonical sampling with the sampling weight as \( 1/\tilde{g}'(E) \), where \( \tilde{g}'(E) \) is the final estimator obtained from the preceding WL or REWL runs. During the simulation, each encountered
structure with lowest energy will be translated into a unique sequence of directions (SoD) containing five elements: Forward (F), Left (L), Right (R), Up (U) and Down (D), and will be stored in a tree data structure for efficient searching and comparison. When a ground-state structure is sampled, its SoD will first be compared with existing ones. Any newly discovered SoD will be added to the database. The whole process terminates when the ground-state degeneracy has “converged” (i.e., no further ground state is discovered for a period of time). Since a SoD uniquely represents a structure (excluding symmetry operations of the underlying simple cubic lattice), this algorithm provides an effective means of identifying and storing different ground-state structures and thus, estimating degeneracies.

4. Results and discussion

![Figure 2](image)

**Figure 2.** Observables from REWL simulations of HP3D46 and H0P3D46 lattice models for Crambin: (top) the densities of states with their ground-state degeneracies (GSD) and (bottom) specific heat ($C_V/N$) of HP3D46, H0P3D46 and semi-flexible H0P3D46. In both figures, error bars smaller than data points are not shown; tuples in the legend indicate the values of $(\epsilon_{HH}, \epsilon_{HO}, \epsilon_{\theta})$.

Crambin is a hydrophobic protein \cite{39} with 46 residues. Its amino acid sequence has been converted to a HP sequence (denoted as HP3D46$^1$) by Lattman et al. \cite{40} and also to a H0P

$^1$ HP3D46: РРНРНРНРРРРНРНРРРРРРРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНРНR

VIII Brazilian Meeting on Simulational Physics IOP Publishing
Journal of Physics: Conference Series 686 (2016) 012001 doi:10.1088/1742-6596/686/1/012001
Figure 3. Top (a, b and c): Three ground-state structures of H0P3D46 with $\epsilon_{HH} = 2$, $\epsilon_{HO} = 1$ and $\epsilon_{\theta} = -0.5$. Hydrophobic and “neutral” monomers are colored in gray and white, respectively, while polar monomers are colored in orange. Bottom (e, f and g): Corresponding structures showing bonds only.
depends upon the details of the model.

Although the peak indicating the folding transition of semi-flexible H0P3D46 is at a slightly higher temperature than for the H0P model without the bond angle term, the specific heat rises more quickly at low temperature (as seen in the inset of Figure 2 (bottom)) due to the smaller energy gap between ground states and first excited states for semi-flexible H0P3D46 (as seen in Figure 2 (top)).

5. Conclusion

Mapping a real protein, Crambin, onto different lattice protein models, the HP model and the H0P model, and using the Replica-Exchange Wang–Landau sampling method, we were able to determine their thermodynamic properties as well as ground-state structures. We found that with the relatively minor generalization of the HP model, i.e. introducing the “neutral” monomer and bond stiffness (semi-flexible H0P model), we were able to reduce the ground-state degeneracies to extremely small values. Moreover, the ground-state structures obtained during the simulation contain features of a “lattice helix” structure which agrees with the experimental results.

Acknowledgments

This work was supported by NSF under Grant No. OCI-0904685; in part by resources and technical expertise from the Georgia Advanced Computing Resource Center, a partnership between the University of Georgias Office of the Vice President for Research and Office of the Vice President for Information Technology; part of the computing resources were provided by the Texas Advanced Computing Center under XSEDE Grant No. PHY130014.

References

[1] Dill K A 1999 *Protein Sci.* 8 1166–80
[2] Kolinski A and Skolnick J 2004 *Polymer* 45 511–524
[3] Dill K A 1985 *Biochemistry* 24 1501–9
[4] Lau K F and Dill K A 1989 *Macromolecules* 22 3986–3997
[5] Irbäck A and Troein C 2002 *J. Biol. Phys.* 28 1–15
[6] Zhang J L and Liu J S 2002 *J. Chem. Phys.* 117 3492
[7] Zhang J, Kou S C and Liu J S 2007 *J. Chem. Phys.* 126 225101
[8] O’Toole E M and Panagiotopoulos A Z 1992 *J. Chem. Phys.* 97 8644
[9] Beutler T C and Dill K A 1996 *Protein Sci.* 5 2037–2043
[10] Grassberger P 1997 *Phys. Rev. E* 56 3682–3693
[11] Frauenkron H, Bastolla U, Gerstner E, Grassberger P and Nadler W 1998 *Phys. Rev. Lett.* 80 3149–3152
[12] Bachmann M and Janke W 2003 *Phys. Rev. Lett.* 91 208105
[13] Hsu H P, Mehra V, Nadler W and Grassberger P 2003 *J. Chem. Phys.* 118 444
[14] Prellberg T and Krawczyk J 2004 *Phys. Rev. Lett.* 92 120602
[15] Tang W and Zhou Q 2012 *Phys. Rev. E* 86 031909
[16] Unger R and Moult J 1993 *J. Mol. Biol.* 231 75
[17] König R and Dandekar T 1999 *BioSystems* 50 17–25
[18] Liang F and Wang W H 2001 *J. Chem. Phys.* 115 3374
[19] Shemygolea A and Hoos H H 2005 *BMC Bioinf.* 6 30
[20] Wüst T and Landau D P 2009 *Phys. Rev. Lett.* 102(17) 178101
[21] Wüst T and Landau D P 2012 *J. Chem. Phys.* 137 064903
[22] Holzgräfe C, Irbäck A and Troein C 2011 *J. Chem. Phys.* 135 195101
[23] Shi G, Vogel T, Wüst T, Li Y W and Landau D P 2014 *Phys. Rev. E* 90 033307
[24] 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
[25] Pattanasiri B, Li Y W, Landau D P, Wüst T and Triampo W 2012 *J. Phys.: Conf. Ser.* 402 012048
[26] Castells V, Yang S and Van Tauvel P 2002 *Phys. Rev. E* 65 031912
[27] Swetnam A and Allen M P 2012 *Phys. Rev. E* 85 062901
[28] Radhakrishna M, Sharma S and Kumar S K 2012 *J. Chem. Phys.* 136 114114
[29] Li Y W, Wüst T and Landau D P 2013 Phys. Rev. E 87 012706
[30] Shi G, Wüst T, Li Y W and Landau D P 2015 J. Phys.: Conf. Ser. 640 012017
[31] Kyte J and Doolittle R F 1982 J. Mol. Biol. 157 105 – 132
[32] Thomas P D and Dill K A 1993 Protein Science 2 2050–2065
[33] Bastolla U and Grassberger P 1997 J. of Stat. Phys 89 1061
[34] Krawczyk J, Owczarek A and Prellberg T 2009 Physica A 388 104–112
[35] Wang F and Landau D P 2001 Phys. Rev. Lett. 86 2050–2053
[36] Wang F and Landau D P 2001 Phys. Rev. E 64 1–16
[37] Vogel T, Li Y W, Wüst T and Landau D P 2013 Phys. Rev. Lett. 110 210603
[38] Vogel T, Li Y W, Wüst T and Landau D P 2014 Phys. Rev. E 90 023302
[39] Schmidt A, Teeter M, Weckert E and Lamzin V S 2011 Acta Crystallographica Section F 67 424–428
[40] Lattman E E, Fiebig K M and Dill K A 1994 Biochemistry 33 6158–6166
[41] Shimada J, Kussell E L and Shakhnovich E I 2001 J. of Mol. Biol. 308 79 – 95