Protein folding of the H0P model: A parallel Wang–Landau study

G Shi¹, T Wüst², Y W Li³ and D P Landau¹

¹Center for Simulational Physics, The University of Georgia, Athens, GA 30602, USA
²Scientific IT Services, ETH Zürich IT Services, 8092 Zürich, Switzerland
³National Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
E-mail: sgjerry@physast.uga.edu

Abstract. We propose a simple modification to the hydrophobic-polar (HP) protein model, by introducing a new type of monomer, “0”, with intermediate hydrophobicity of some amino acids between H and P. With the replica-exchange Wang–Landau sampling method, we investigate some widely studied HP sequences as well as their H0P counterparts and observe that the H0P sequences exhibit dramatically reduced ground state degeneracy and more significant transition signals at low temperature for some thermodynamic properties, such as the specific heat.

1. Introduction

For decades, scientists have endeavored to understand the secret code of protein folding, which is believed to be essential in determining the shape of the native state and thus the biological function of a protein. Beside fundamental research, the motivation for studying the process of protein folding also comes from the large number of diseases in humans resulting from protein misfolding, e.g., Alzheimer’s, Parkinson’s, etc. [1]. Despite the astronomical number of possible configurations, a protein folds into its functional, three-dimensional structure (native state) amazingly fast; nevertheless, the mechanism behind folding is still unclear [2].

Generic, coarse-grained protein models have been developed to help understand the properties of protein folding [3, 4]. Among these, the hydrophobic-polar (HP) lattice protein model [5, 6] is arguably the simplest, and thus has been called the “Ising model” for protein folding [7]. The HP model attracts researchers from various disciplines because of its simplicity yet challenging computational complexity. It has been proven that finding the lowest energy structure of a given sequence is an NP-complete problem [8]; however, many folding algorithms and Monte Carlo methods have been developed for approaching this problem. Some representative examples include, but are not limited to, chain-growth methods [9, 10], e.g. the pruned-enriched Rosenbluth method (PERM) and its variants [11, 12, 13, 14, 15], sequential importance sampling [16], fragment regrowth Monte Carlo [17], multidomain sampler [18], genetic algorithm [19, 20], evolutionary Monte Carlo [21], ant colony models [22] and Wang–Landau sampling [23, 24].

The HP model has been used in several problems of biological interest, such as protein surface adsorption [25, 26, 27, 28], protein folding in membranes [29] or confined environments [30, 31], and the effects of sequence mutations [32, 33]. However, simplification into only H and P...
monomers yields large ground state degeneracy which stands in contrast to the generally unique native state of natural proteins. More advanced models such as the HPNX model [34, 35], which considers not only the hydrophobicity but also the charges of amino acids, has been shown to be effective in reducing degeneracy. From a different perspective, we propose a simple modification to the original HP model, by introducing a new type of “neutral” monomer, “0”, i.e., neither hydrophobic nor polar, thus rendering the model more realistic without significantly increasing the difficulties of sampling.

2. Model and methods

2.1. H0P lattice protein model

The HP lattice model classifies amino acids into just two types, hydrophobic (H) and polar (P), and the structure of a protein is simulated by a chain of connected beads, each of which represents a single monomer, on a simple cubic lattice. The only interaction considered in the HP model is an effective monomer-monomer coupling $\epsilon_{HH}$ between non-bonded, nearest-neighbor H monomers. However, most HP sequences suffer from the drawback of quite large ground-state degeneracy.

Since, in reality, different amino acids possess different levels of hydrophobicity [36], simplification into only two types might be insufficient to characterize the features of the hydrophobic interaction, which is believed to be the driving force of protein folding and tertiary structure formation. Therefore, we introduce a new level of “neutral” monomers, “0”, which neither favors nor dislikes water as much as H or P monomers do. In the H0P model, two different non-bonded interactions are considered: $\epsilon_{HH}$ and $\epsilon_{H0}$, where $\epsilon_{H0}$ denotes the interaction strength between the H and the “0” monomers, and $\epsilon_{HH} > \epsilon_{H0}$. The Hamiltonian is then given by

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

(1)

where $n_{HH}$ is the number of non-bonded HH contacts ($n_{H0}$ has corresponding meaning). We used $\epsilon_{HH} = 2$ and $\epsilon_{H0} = 1$ for this work, but other interactions involving “0” monomers could be easily added.

2.2. Replica-exchange Wang–Landau sampling

We have used a recently developed generic, parallel Wang–Landau sampling scheme, namely Replica-exchange Wang–Landau sampling (REWL) [37, 38]. In this parallel scheme, the energy range of the system is split into multiple, overlapping windows, each of which is simulated by independent processes (random walkers, running serial Wang–Landau (WL) sampling [39, 40]). The replica exchanges between overlapping windows are attempted during the simulation at a fixed time interval, such that each replica can travel through the entire energy space of interest.

WL sampling is a serial Monte Carlo simulation method that has proven to be powerful for studying problems with complex free energy landscapes. The ultimate goal of WL sampling is to estimate the density of states $g(E)$, using a random walker in an energy range $[E_{\text{min}}, E_{\text{max}}]$. During the simulation, a MC trial move that changes the system from configuration A (with energy $E_A$) to configuration B (with energy $E_B$) will be accepted with probability

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

(2)

where $g'(E)$ is an instantaneous estimator for the density of states. Upon acceptance, this estimator is updated via $g'(E_B) \rightarrow f \times g'(E_B)$, where $f$ is a modification factor, and the histogram of visited energies is increased by $H(E_B) \rightarrow H(E_B) + 1$. If the trial move was rejected, $g'(E_A)$ and $H(E_A)$ are updated instead in the same way. The modification factor will
be reduced if the histogram is sufficiently "flat" and at the same time, all the histogram entries are reset to zero. The above procedures will be performed iteratively, 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 $f_{\text{final}} = \exp(10^{-8})$. Initially, $g'(E)$ is set to 1 and we use 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. Multiple, independent runs are performed to test for systematic errors and to determine statistical errors. As Monte Carlo trial moves, we applied pull and bond-rebridging moves which have proven very efficient together with WL sampling [23, 24].

In order to study the ground-state properties of lattice proteins, we proposed a ground-state sampling method [33] to estimate the ground-state degeneracy and collect the ground-state structures. The basic idea behind this method is to perform a multicanonical sampling using $1/\tilde{g}(E)$ as the sampling weight, where $\tilde{g}(E)$ is the final estimator obtained from the preceding WL or REWL run. During the process, each encountered ground-state structure, i.e. the structure obtains the lowest energy found during the WL or REWL run, will be converted 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. When a ground-state structure is generated, its SoD will first be compared with existing ones. Any newly discovered structure will be added to the database. The simulation 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 ground-state structure with symmetries of a simple cubic lattice taken into account, this method provides an effective means of identifying and storing different ground-state structures and determine the degeneracy.

3. Results and discussion

3.1. Randomized Sequences

Our first step is to test the H0P model on one of the widely studied 48mers (“Harvard sequences”), 48.1 [41], which has been shown to have enormous ground-state degeneracy in previous studies [33, 42]. This sequence was originally designed for testing of folding algorithms and, thus, there is not a defined way to map it into a H0P sequence. Therefore, we randomly generated four H0P sequences based on 48.1 by converting each monomer (either H or P) to the new type 0 with a probability of one third, as shown in Table 1. After performing REWL sampling as well as the ground-state degeneracy estimation procedure, we list the ground-state degeneracies for all four sequences in Table 1. The ground-state degeneracies are quite limited for all four H0P sequences, and are orders of magnitude less than that for the equivalent HP 48mer.

Table 1. Estimation of ground-state degeneracy of four randomly generated H0P sequences with 48 monomers. Sequences with converged degeneracy do not have statistical errors.

| ID        | sequence                                      | ground-state degeneracy |
|-----------|-----------------------------------------------|-------------------------|
| 48.1      | HPHP_{2}P_{4}H_{3}P_{2}H_{2}P_{3}H_{3}PH_{3}P_{2}H_{2}P_{3}HP_{4}H_{2} | 1.03(4) × 10^{7} [33] |
| 48.1-1    | 0_{5}HP_{0}H_{2}O_{0}PH_{0}(0HP)_{2}H_{0}(0P)_{2}H(H002P_{2}P_{0})_{2}0P_{3}O_{3}H_{2} | 269 |
| 48.1-2    | HP_{0}HP_{2}(H0)_{2}P_{0}O_{0}PH_{2}(PH)_{2}0_{2}H_{0}P_{0}H_{2}P_{0}P_{0}O_{2}P_{0}O_{2}H_{0} | 1561 |
| 48.1-3    | HPH_{0}P_{0}O_{0}H_{0}P_{0}H_{0}O_{0}P_{0}O_{2}(HP)_{2}P_{0}O_{2}(0P)_{2}O_{2}P_{0}P_{0}O_{2}(P)_{2}O_{2}H_{0} | 34 |
| 48.1-4    | HPH_{2}P(0H)_{2}O_{2}P_{0}O_{2}P_{0}H_{0}O_{2}O_{2}H_{0}P_{0}O_{2}(0P)_{2}O_{2}P_{0}O_{2}(0P)_{2}O_{2}H_{0} | 3695 |
3.2. A Real Protein

Table 2. Amino acid sequence of Crambin, as well as its HP and H0P mappings.

| ID       | sequence                                      |
|----------|-----------------------------------------------|
| Crambin  | TTCCPSIVARSNFVCPRTPEALCATYTGCHPGATCPGDYAN    |
| HP3D46 [43] | PPHHPPHHHPPHPHPPPHHHPPHHHHPPHPHPPPHHPHPPPHHP |
| H0P3D46  | 00HH00HH00P0PHPHHPH0000PHH0000HHHH0000H00P00P |

Figure 1. Physical observables from WL/REWL simulations of HP3D46 and H0P3D46 lattice models for Crambin: (a) the densities of states with their ground-state degeneracies (GSD) and (b) the specific heat of HP3D46 and H0P3D46 sequences. In both figures, error bars smaller than data points are not shown.

As another example, we choose a real protein, Crambin, which has 46 residues and has been converted into a HP sequence by Lattman et al. [43]. Based on the hydrophobic index table
in [36], we defined Ile, Val, Leu, Phe and Cys as “H”, Met, Ala, Gly, Thr, Trp, Ser, Tyr and Pro as “0”, and the rest as “P”. The resulting H0P sequence (H0P3D46) is shown in Table 2, along with the residue sequence of Crambin and its (simpler) HP mapping (HP3D46).

The densities of states, as the direct results of our simulations, are shown in Fig. 1 (a), in which we also label the ground-state degeneracies. Along with a lower ground-state energy, the H0P model also has a significantly lower ground-state degeneracy than the HP model. For comparing the ground-state structures, we calculate the contact maps for ground states of HP3D46 and H0P3D46, by assigning +1 to energy-contributing contacts (HH contacts for HP model; HH and H0 contacts for H0P model) and −1 to all the other (zero-energy) contacts. These contact maps are averaged to obtain the contact map density and is shown in Fig. 2. From this figure, we could tell that the H0P model compared to the HP model, has many fewer zero-energy contacts, which also explains the lower ground-state degeneracy. The specific heat ($C_V/N$) in Fig. 1 (b), is quite different for HP3D46 and H0P3D46. While H0P3D46 shows very clear signals of a two-step folding process as observed in all-atom simulations [44], the coarseness of the lattice model does not allow formation of the clear, connected two-helix structure found by experiments (see [45] and the references therein).

4. Conclusion
In conclusion, we improved the original HP model by considering more complex hydrophobicity levels and thus introduced a new type of monomer “0”. With replica-exchange Wang–Landau sampling, we investigated some widely studied HP sequences and their H0P counterparts. As expected, H0P model shows more pronounced transition signals (e.g. $C_V/N$) in the low temperature region, as a result of its lower ground-state degeneracy.

Acknowledgments
This work was supported by NSF under Grant No. OCI-0904685. Y.W.L. was sponsored by the Oak Ridge Leadership Computing Facility, the Office of Advanced Scientific Computing Research; U.S. Department of Energy. Part of the work was carried out at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC under Contract No. DE-AC05-000R22725. Computing resources were partially provided by the Georgia Advanced Computing Resource Center and the Texas Advanced Computing Center under XSEDE Grant No. PHY130014.
References
[1] Chaudhuri T K and Paul S 2006 The FEBS J. 273 1331–49
[2] Dill K A and MacCallum J L 2012 Science 338 1042–6
[3] Dill K A 1999 Protein Sci. 8 1166–80
[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] Li Y W, Wüst T and Landau D P 2011 Comput. Phys. Commun. 182 1896–1899
[8] Irbäck A and Troein C 2002 J. Biol. Phys. 28 1–15
[9] O’Toole E M and Panagiotopoulos A Z 1992 J. Chem. Phys. 97 8644
[10] Beutler T C and Dill K A 1996 Protein Sci. 5 2037–2043
[11] Grassberger P 1997 Phys. Rev. E 56 3682–3693
[12] Frauenkron H, Bastolla U, Gerstner E, Grassberger P and Nadler W 1998 Phys. Rev. Lett. 80 3149–3152
[13] Bachmann M and Janke W 2003 Phys. Rev. Lett. 91 208105
[14] Hsu H P, Mehra V, Nadler W and Grassberger P 2003 J. Chem. Phys. 118 444
[15] Prellberg T and Krawczyk J 2004 Phys. Rev. Lett. 92 120602
[16] Zhang J L and Liu J S 2003 J. Chem. Phys. 117 3492
[17] Zhang J, Kou S C and Liu J S 2007 J. Chem. Phys. 126 225101
[18] Tang W and Zhou Q 2012 Phys. Rev. E 86 031909
[19] Unger R and Moult J J 1993 J. Mol. Biol. 231 75
[20] König R and Dandekar T 1999 BioSystems 50 17–25
[21] Liang F and Wong W H 2001 J. Chem. Phys. 115 3374
[22] Shemygelska A and Hoos H H 2005 BMC Bioinf. 6 30
[23] Wüst T and Landau D P 2009 Phys. Rev. Lett. 102(17) 178101
[24] Wüst T and Landau D P 2012 J. Chem. Phys. 137 064903
[25] Castells V, Yang S and Van Tassel P 2002 Phys. Rev. E 65 031912
[26] Swetnam A and Allen M P 2012 Phys. Rev. E 85 062901
[27] Radhakrishna M, Sharma S and Kumar S K 2012 J. Chem. Phys. 136 114114
[28] Li Y W, Wüst T and Landau D P 2013 Phys. Rev. E 87 012706
[29] Bonaccini R and Seno F 1999 Phys. Rev. E 60 7290–8
[30] 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
[31] Pattanasiri B, Li Y W, Landau D P, Wüst T and Triampo W 2012 J. Phys.: Conf. Ser. 402 012048
[32] Holzgräfe C, Irbäck A and Troein C 2011 J. Chem. Phys. 135 195101
[33] Shi G, Vogel T, Wüst T, Li Y W and Landau D P 2014 Phys. Rev. E 90 033307
[34] Bornberg-Bauer E 1997 Proceedings of the First Annual International Conference on Computational Molecular Biology RECOMB ‘97 (New York, NY, USA: ACM) pp 47–55 ISBN 0-89791-882-7
[35] Hoque T, Chetty M and Sattar A 2009 J. Comput. Biol 16 85–103
[36] Kyte J and Doolittle R F 1982 J. Mol. Biol. 157 105 – 132
[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] Wang F and Landau D P 2001 Phys. Rev. Lett. 86 2050–2053
[40] Wang F and Landau D P 2001 Phys. Rev. E 64 1–16
[41] Yue K, Fiebig K M, Thomas P D, Chan H S, Shakhnovich E I and Dill K A 1995 PNAS 92 325–329
[42] Bachmann M and Janke W 2004 J. Chem. Phys. 120 6779–91
[43] Lattman E E, Fiebig K M and Dill K A 1994 Biochemistry 33 6158–6166
[44] Shimada J, Kussell E L and Shakhnovich E I 2001 J. of Mol. Biol. 308 79–95
[45] Schmidt A, Teeter M, Weckert E and Lamzin V S 2011 Acta Crystallographica Section F 67 424–428