Replica Exchange Wang–Landau Simulation of Lattice Protein Folding Funnels

Guangjie Shi[1], **Thomas Wüst**[2] **and David P. Landau**[1]

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

E-mail: `sgjerry@physast.uga.edu`

Abstract. The resolution of Levinthal's paradox concerning the ability of proteins to fold rapidly postulates the existence of a rough "folding funnel" in free energy space that guides the protein to its lowest free energy, native state. To study the folding of the protein ribonuclease A we have mapped it onto a 124 monomer, coarse-grained HP lattice model and onto an H0P model that also includes "neutral" 0-mers in addition to the hydrophobic H-mers and polar P-mers. Using Replica Exchange Wang-Landau sampling, we determined the density of states, g(E), which we then used to calculate the free energy of the protein vs end-to-end distance as a function of temperature. At low temperature the HP model shows a rather shallow and flat free energy minimum, while the H0P model maintains a rough free energy funnel. Unlike the common, schematic figures, we find an asymmetric folding funnel that also changes shape substantially as the temperature decreases. Even the location of the free energy minimum shifts as the temperature decreases.

1. Introduction

Understanding protein folding remains a Grand Challenge problem of modern science. The protein folding funnel concept [1, 2], which is believed to be the explanation of the protein folding problem, has always been portrayed schematically as a function of some unknown reaction coordinate (as seen in Fig. 1) and has never actually been observed. However, due to the complexity of the problem, investigation is only possible through the combinations of coarse-grained models and advanced computational methods. The hydrophobic-polar (HP) lattice protein model [3, 4] is arguably the simplest protein model, yet it has contributed to the understanding of many problems of biological interest [5]-[6] An extension of the HP protein model, the semi-flexible H0P lattice protein model [7], introduced some simple modifications which render the model more realistic with significantly reduced ground-state degeneracy [8].

Even though the HP model is highly simplified, the problem of finding the ground-state structure of a given HP sequence is extremely difficult [9] and ground states are often highly degenerate. Consequently, the HP model represents a significant challenge at the interface between statistical physics and biophysics. The HP model has thus become a testing ground for many advanced algorithms: sequential importance sampling [10], fragment regrowth Monte Carlo [11], chain-growth methods [12, 13], e.g. the pruned-enriched Rosenbluth method (PERM) and its variants [14, 15, 16, 17, 18], multidomain sampler [19], genetic algorithm [20, 21], evolutionary Monte Carlo [22], ant colony models [23] and Wang–Landau sampling [24, 25]. Here we have applied an efficient, parallel version of the Wang–Landau method, i.e. Replica

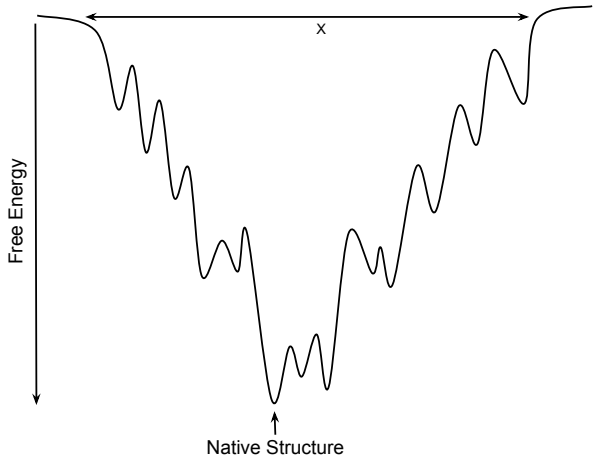


Figure 1. Schematic view of a rough protein folding funnel.

Exchange Wang–Landau (REWL) sampling to both the HP model and the semi-flexible H0P model to uncover protein folding funnels.

2. Model and methods

2.1. HP and semi-flexible H0P lattice protein models

Based on the properties of their side chains, amino acids are classified into hydrophobic (H) and polar (P) in the HP lattice protein model [3, 4]. In this model, a chain of connected monomers represents the amino acid sequence residing on a simple cubic lattice. An effective monomer-monomer coupling $\epsilon_{\rm HH}$ between non-bonded, nearest-neighbor H monomers is used for characterizing the driving force of protein folding: the hydrophobic interaction:

$$\mathcal{H} = -\epsilon_{\rm HH} n_{\rm HH}, \tag{1}$$

where $n_{\rm HH}$ is the number of non-bonded HH contacts.

An improved version of the HP model, i.e. a semi-flexible H0P lattice protein model, introduces a "neutral" type monomer ("0") representing amino acids that are neither hydrophobic nor polar. This model also considers the stiffness of angles formed by bonds connecting monomers, and the general Hamiltonian of the semi-flexible H0P model can be written as

$$\mathcal{H} = -\epsilon_{\rm HH} n_{\rm HH} - \epsilon_{\rm H0} n_{\rm H0} - \epsilon_\theta n_\theta, \tag{2}$$

where $n_{\rm HH}$ is the number of non-bonded HH contacts ($n_{\rm H0}$ has corresponding meaning), and n_θ represents the number of angles existing in the protein structure (see Fig.2). We used $\epsilon_{\rm HH} = 1$, $\epsilon_{\rm H0} = 0.5$ and $\epsilon_\theta = -0.25$ for this work, but other interaction values or interactions involving "0" monomers (or more distant neighbors) could be easily added. Our intent, however, was to add only as many features as are needed to render the original HP model more realistic while still keeping the model as simple as possible.

2.2. Wang–Landau sampling

Wang–Landau (WL) sampling [26, 27] is an iterative Monte Carlo (MC) simulation method, which performs a random walk over the energy space for estimating the density of states of the system under investigation. This method, using pull moves and bond-rebridging moves to

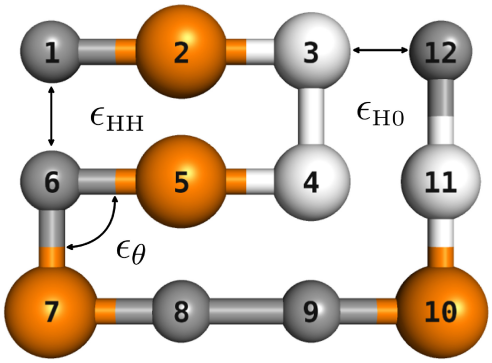


Figure 2. A schematic example of a semi-flexible H0P model. The interaction between monomers 1 and 6 is ϵ_{HH}, and that between monomers 3 and 12 is ϵ_{H0}. The angle constituted by monomers 5, 6 and 7 contributes ϵ_{θ} energy. In this particular 2-dimensional structure, $n_{\mathrm{HH}} = 1$, $n_{\mathrm{H0}} = 2$ and $n_{\theta} = 5$. Hydrophobic and "neutral" monomers are colored dark gray and white, respectively, while polar monomers are colored orange. An HP model results if no 0-mers are present and the angle energy is zero. Both models are actually simulated in three dimensions.

generate new, trial states, has been shown to be powerful and efficient in both finding ground state structures and determining density of states ($g(E)$) [24, 25]. During the simulation, a MC trial move, changing the system from energy E_A to energy E_B, will be accepted with probability

$$P(A \to B) = \min\left\{1, \frac{g'(E_A)}{g'(E_B)}\right\}, \tag{3}$$

where $g'(E)$ is an instantaneous estimator for the density of states. This estimator is updated through $g'(E_B) \to f \times g'(E_B)$ at each step, where f is a modification factor. The simulation maintains a histogram of visited energies, which is increased by $H(E_B) \to H(E_B) + 1$. 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 simulation continues until the modification factor is less than some predefined threshold value f_{final}. In this work, the initial modification factor is set to $f_{init} = e^1$ and decreased via $f \to \sqrt{f}$, until reaching $f_{\mathrm{final}} = 1 \times 10^{-8}$. 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 at each iteration.

2.3. Replica Exchange Wang–Landau sampling

For determining $g(E)$ of systems with the size of interest in this work, it was necessary to employ Replica Exchange Wang–Landau sampling (REWL) [28, 29], a generic, parallel Wang–Landau sampling scheme. As shown in Fig. 3, the energy range of the system is split into multiple, overlapping windows, each of which is simulated by independent processes (random walkers, running serial WL sampling). 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. The simulation terminates only when all random walkers have satisfied the termination criteria independently.

3. Results and discussion

REWL sampling was performed for both the HP sequence (denoted as HP124) [30] and the H0P sequence (denoted as H0P124) based on the mapping of ribonuclease A. (See ref. [7] for more

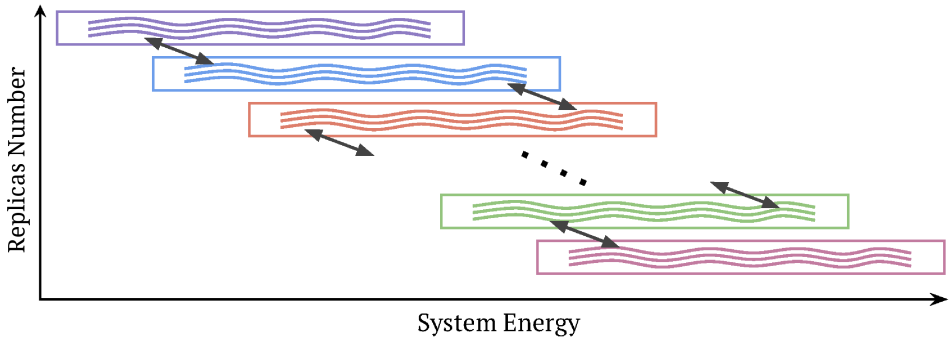


Figure 3. Illustration of the general framework of Replica Exchange Wang–Landau sampling. The system energy range is split into multiple, overlapping energy windows. Each window employs multiple independent processes (or walkers) running serial Wang–Landau sampling. At fixed time intervals, the replica exchange procedure is performed between two neighboring energy windows.

about the mapping.) The density of states was determined to high precision for both models and used to estimate the partition function and both thermodynamic and structural properties. If Q represents a structural quantity which can be used as a reaction coordinate, we can calculate the free energy with respect to temperature and Q based on the joint density of states $g(E, Q)$ as:

$$F\,(T, Q) = -k_B\,T\,\ln Z(T, Q), \tag{4}$$

where $Z(T, Q)$ is the partition function based on temperature and Q:

$$Z\,(T, Q) = \sum_E g(E, Q)\,e^{-E/k_B T}. \tag{5}$$

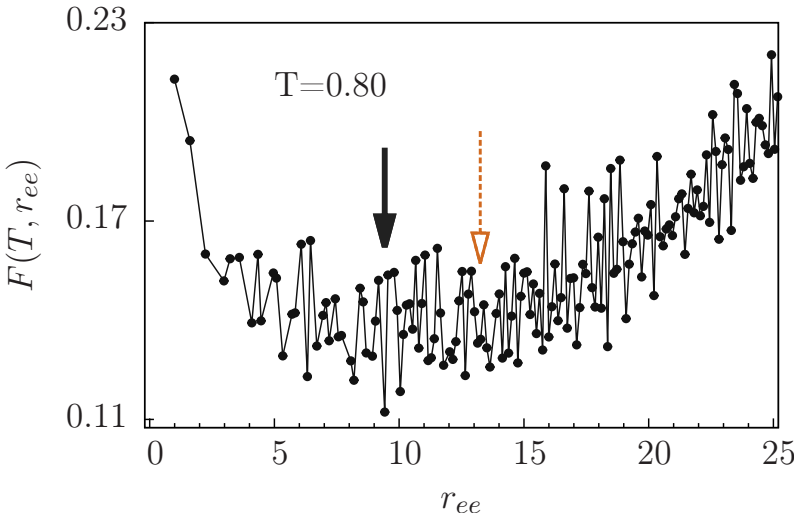
After initial exploration, we found that the end-to-end distance:

$$r_{ee} = \,|\vec{r}_N - \vec{r}_1|$$

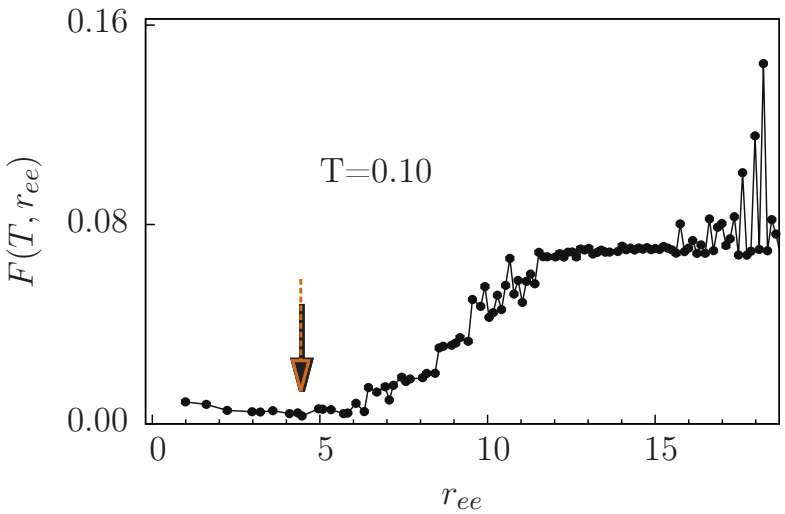
was apparently a good candidate for representing the folding funnel, where N is the number of monomers in the chain; $\vec{r}_i$ represents the position of the ith monomer.

In Fig. 4 we present the normalized free energy vs end-to-end distance at two different temperatures for the HP124 lattice protein. The quality of the results was quite high and the statistical errors in the figure are smaller than the size of the symbols. The state, and associated end-to-end distance, with lowest free energy is indicated by a black arrow, while the mean r_{ee} is marked by an orange arrow. At high temperature (as seen in Fig. 4(a)), the free energy shows a shallow, "symmetric" landscape which contains many local maxima and minima. The state with lowest free energy has a different end-to-end distance than the mean end-to-end distance, and there are many minima and maxima between the two states. Upon lowering the temperature we found the free energy curve changes shape and becomes skewed toward the region with low r_{ee} values (as seen in Fig. 4(b)). Moreover, unlike the static schematic portrayals of the protein folding funnel which always has a fixed minimum, the lowest free energy position shifts with the change in temperature, showing the dynamic nature of the folding funnel. Finally, at quite low temperatures as seen in Fig. 4(b), the free energy landscape of HP124 becomes relatively flat near the minimum, and the system can easily move between states with different end-to-end

distances. At low temperature, the state with the lowest free energy is then also the state with the mean end-to-end distance. For greater end-to-end distances, the free energy curve assumes a rather odd shape with a flat plateau region.



(a)



(b)

Figure 4. Normalized free energy vs end-to-end distance at (a) $T = 0.8$ and (b) $T = 0.1$ for the HP124 lattice protein. Black arrows indicate the lowest free energy at each temperature (T), orange arrows point to the mean end-to-end distance at that temperature. Error bars are smaller than the data points.

The normalized free energy vs end-to-end distance at two different temperatures for the H0P124 lattice protein shows significant differences from its HP counterpart. Results are shown in Fig. 5. The state with lowest free energy is indicated by a black arrow, while the mean r_{ee} is marked by an orange arrow. At high temperatures (as seen in Fig. 5 (a)) we observed a rugged free energy landscape which is similar to that for HP124, but as the temperature is lowered, substantial differences become apparent. The free energy landscape remains rough but skews towards the region with low r_{ee} values, and the position of the lowest free energy also shifts

with temperature. However, at low temperatures, unlike the free energy landscape of HP124, that of H0P124 remains rough even near the bottom (as seen in Fig. 5 (b)) and does not show the curious, plateau region. The inset shows clearly that significant free energy barriers must be overcome to move between free energy minima and the protein can be easily trapped in a metastable state if traditional simulation methods are used.

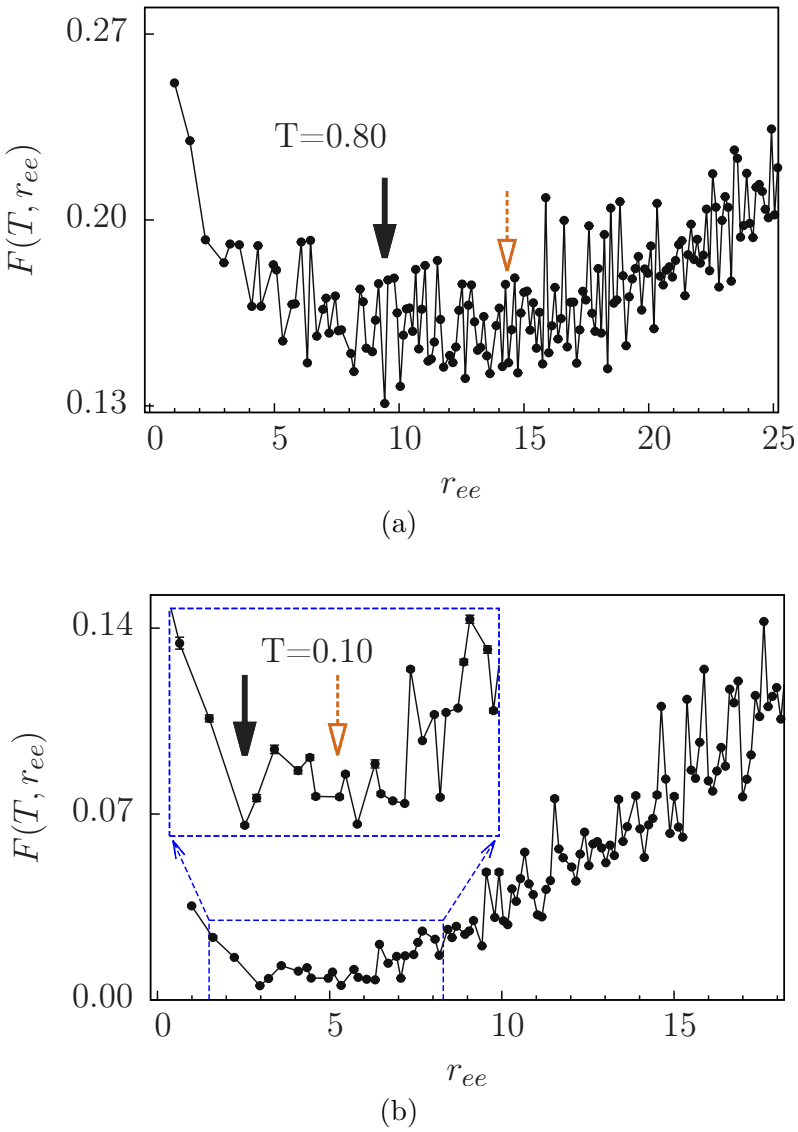


Figure 5. Normalized free energy vs end-to-end distance at (a) $T = 0.8$ and (b) $T = 0.1$ for the H0P124 lattice protein. Black arrows indicate the lowest free energy at each temperature (T), orange arrows point to the mean end-to-end distance at that temperature. Error bars are smaller than the data points.

4. Conclusion

In summary, we investigated the free energy landscapes for two lattice protein models that are mapped from the protein ribonuclease A. Using the end-to-end distance as the reaction coordinate, we mapped out the folding funnels for these two model proteins. The HP model has

a relatively shallow free energy minimum at low temperature, while the H0P model develops a clear, rough free energy funnel, even near the minimum, at all temperatures. In both cases, we find an asymmetric folding funnel that changes shape substantially as the temperature changes. Unlike the schematic figures in the literature, we even observe that the location of the free energy minimum shifts with the change of temperature. Even though both model proteins are highly simplified descriptions of a real one, neither the mapping nor the interactions were tuned to produce a special free energy structure. Moreover, with only a few features added to the HP model, the H0P model already captures some of the essential thermodynamic and folding behavior of real proteins (e.g. rough free energy funnel). Therefore, we believe that the general characteristics of the folding funnels discovered in this work might well persist in a more realistic description of protein folding.

Acknowledgments
This work was supported in part by NSF under Grant No. OCI-0904685. 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] Dill K A and Chan H S 1997 *Nat Struct Mol Biol* **4** 10–19
[2] Onuchic J N and Wolynes P G 2004 *Curr. Opin. Struct. Biol.* **14** 70 – 75 ISSN 0959-440X
[3] Dill K A 1985 *Biochemistry* **24** 1501–9
[4] Lau K F and Dill K A 1989 *Macromolecules* **22** 3986–3997
[5] Castells V, Yang S and Van Tassel P 2002 *Phys. Rev. E* **65** 031912
[6] Pattanasiri B, Li Y W, Landau D P, Wüst T and Triampo W 2012 *J. Phys.: Conf. Ser.* **402** 012048
[7] Shi G, Wüst T, Li Y and Landau D 2015 *J. Phys.: Conf. Ser.* **640** 012017
[8] Shi G, Wüst T, Li Y W and Landau D P 2016 *J. Phys.: Conf. Ser.* **686** 012001
[9] Irbäck A and Troein C 2002 *J. Biol. Phys.* **28** 1–15
[10] Zhang J L and Liu J S 2002 *J. Chem. Phys.* **117** 3492
[11] Zhang J, Kou S C and Liu J S 2007 *J. Chem. Phys.* **126** 225101
[12] O'Toole E M and Panagiotopoulos A Z 1992 *J. Chem. Phys.* **97** 8644
[13] Beutler T C and Dill K A 1996 *Protein Sci.* **5** 2037–2043
[14] Grassberger P 1997 *Phys. Rev. E* **56** 3682–3693
[15] Frauenkron H, Bastolla U, Gerstner E, Grassberger P and Nadler W 1998 *Phys. Rev. Lett.* **80** 3149–3152
[16] Bachmann M and Janke W 2003 *Phys. Rev. Lett.* **91** 208105
[17] Hsu H P, Mehra V, Nadler W and Grassberger P 2003 *J. Chem. Phys.* **118** 444
[18] Prellberg T and Krawczyk J 2004 *Phys. Rev. Lett.* **92** 120602
[19] Tang W and Zhou Q 2012 *Phys. Rev. E* **86** 031909
[20] Unger R and Moult J 1993 *J. Mol. Biol.* **231** 75
[21] König R and Dandekar T 1999 *BioSystems* **50** 17–25
[22] Liang F and Wong W H 2001 *J. Chem. Phys.* **115** 3374
[23] Shmygelska A and Hoos H H 2005 *BMC Bioinf.* **6** 30
[24] Wüst T and Landau D P 2009 *Phys. Rev. Lett.* **102**(17) 178101
[25] Wüst T and Landau D P 2012 *J. Chem. Phys.* **137** 064903
[26] Wang F and Landau D P 2001 *Phys. Rev. Lett.* **86** 2050–2053
[27] Wang F and Landau D P 2001 *Phys. Rev. E* **64** 1–16
[28] Vogel T, Li Y W, Wüst T and Landau D P 2013 *Phys. Rev. Lett.* **110** 210603
[29] Vogel T, Li Y W, Wüst T and Landau D P 2014 *Phys. Rev. E* **90** 023302
[30] Lattman E E, Fiebig K M and Dill K A 1994 *Biochemistry* **33** 6158–6166