PHYSICAL REVIEW E **90**, 033307 (2014)

Effect of single-site mutations on hydrophobic-polar lattice proteins

Guangjie Shi,[1,*] Thomas Vogel,[2] Thomas Wüst,[3] Ying Wai Li,[4] and David P. Landau[1]
[1]*Center for Simulational Physics, The University of Georgia, Athens, Georgia 30602, USA*
[2]*Theoretical Division (T-1), Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[3]*Scientific IT Services, ETH Zürich IT Services, 8092 Zürich, Switzerland*
[4]*National Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA*

(Received 10 April 2014; published 15 September 2014)

We developed a heuristic method for determining the ground-state degeneracy of hydrophobic-polar (HP) lattice proteins, based on Wang-Landau and multicanonical sampling. It is applied during comprehensive studies of single-site mutations in specific HP proteins with different sequences. The effects in which we are interested include structural changes in ground states, changes of ground-state energy, degeneracy, and thermodynamic properties of the system. With respect to mutations, both extremely sensitive and insensitive positions in the HP sequence have been found. That is, ground-state energies and degeneracies, as well as other thermodynamic and structural quantities, may be either largely unaffected or may change significantly due to mutation.

DOI: 10.1103/PhysRevE.90.033307 PACS number(s): 05.10.−a, 87.14.et, 87.15.ak, 87.15.Qt

I. INTRODUCTION

The protein folding problem has been studied for more than 50 years, but much remains to be learned. One of the fundamental remaining questions is [1] "How is the three-dimensional (3D) native structure of a protein determined by the physicochemical properties that are encoded in its one-dimensional amino acid sequence?" Scientists used to believe that any two naturally occurring proteins with a 40% or higher sequence identity would possess the same fold [2]. However, experiments discovered that proteins such as *Pfl6* and *Xfaso 1* with high sequence similarity end up with different folds [3]. Furthermore, Alexander *et al.* successfully designed two proteins that share 88% of their sequence but fold into totally different tertiary structures [4]. More generally, it was recently experimentally confirmed that small local differences can lead to large changes in the global organization of amino acid sequences [5]. To understand these phenomena better, a first step is to investigate how the change of a single amino acid affects higher order structure and functions of a protein. Experimental studies of single amino acid substitutions on *lac* repressor [6], for example, showed that proteins often keep phenotypically silent for about 50% of such substitutions. However, very sensitive positions in the amino acid sequence also exist [7].

The understanding of the protein folding problem has been enhanced through studies of generic, coarse-grained models [8,9]. The simplest one is the hydrophobic-polar (HP) lattice model [10,11], which has been used in several problems of biological interest, such as surface adsorption [12–16] and protein folding in membranes [17] or confined environments [18,19]. Despite the simplicity of the HP model, finding the lowest energy structure of a given sequence is an NP-complete problem [20]. For long sequences (chain length $\gtrsim 30$), enumeration methods (see, e.g., Refs. [21–23]) are not accessible. However, different folding algorithms and Monte Carlo methods have been developed for approaching this problem. Examples include, but are not limited to, constraint-based approaches [24], chain-growth methods [25,26], in particular the pruned-enriched Rosenbluth method (PERM) and its variants [27–31], sequential importance sampling [32], fragment regrowth Monte Carlo [33], multidomain sampler [34], genetic algorithms [35,36], evolutionary Monte Carlo [37], and ant colony models [38]. Among those, the sampling method developed by Wang and Landau [39] has been shown to be powerful and highly precise in simulating proteins and polymers [40,41].

The HP model has been found to have similar mutational properties compared to real proteins [42,43]. A recent study on two-dimensional (2D) HP proteins with chain lengths $\leqslant 30$ shows a single-mutation-induced fold switching [44], which has also been discovered in experimental studies [4,45–50]. The main focus of the present article is on the study of the effect of single-site substitution mutations on multiple HP sequences as a way of systematically approaching some of the questions introduced before on a very fundamental level. Therefore, we also introduce a technique for independently estimating the complete density of states, including the ground-state degeneracy, during a generalized-ensemble simulation for which the simulation weights are determined using Wang-Landau sampling. A key to the efficiency of the approach is the encoding of a three-dimensional HP-protein configuration uniquely into a serial direction sequence. This enables us to store and access the information of all visited structures efficiently during the simulation. In particular, the analysis of ground-state structures becomes feasible and can be carried out conveniently. In Sec. II, we describe the model and method, results are presented in Sec. III, and we conclude in Sec. IV.

II. MODEL AND METHODS

A. The HP lattice model

The HP lattice model is a coarse-grained protein model which classifies amino acids into just two types, hydrophobic (H) and polar (P). Each amino acid in this model is represented as a single monomer on a simple cubic lattice. The hydrophobic interaction, as the key driving force of protein folding and tertiary structure formation [51,52], is characterized by an effective monomer-monomer coupling ϵ_{HH} between nonbonded nearest-neighbor H monomers. The Hamiltonian is

[*] Corresponding author: sgjerry@physast.uga.edu

©2014 American Physical Society

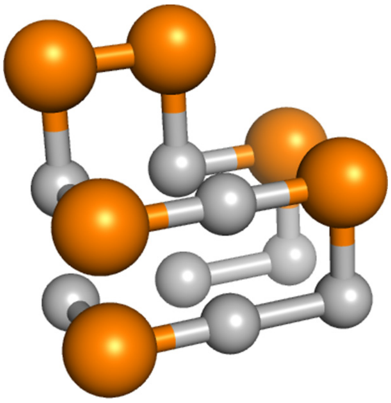

FIG. 1. (Color online) A 3D structure of a HP protein with eight hydrophobic (H: small, silver beads) and six polar (P: large, orange beads) residues.

given by

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

where n_{HH} is the number of nonbonded HH contacts. Hence, the ground state, i.e., lowest-energy state, of a HP protein is the state with a maximum number of n_{HH}. See Fig. 1 for a sample visualization of a small HP ground-state structure with $n_{\mathrm{HH}} = 8$.

B. Wang-Landau sampling and trial moves

Wang-Landau sampling [39] is a Monte Carlo (MC) method which aims to estimate the density of states $g(E)$ while ideally performing a random walk in energy (E) space. The acceptance probability for a MC trial move that changes the system from configuration A (with energy E_A) to configuration B (with energy E_B) is given by

$$P(A \to B) = \min\left\{1, \frac{g'(E_A)}{g'(E_B)}\right\}. \tag{2}$$

Upon acceptance, the estimator $g'(E)$ for the density of states is updated via $g'(E_B) \to f \times g'(E_B)$, where f is a modification factor, and the histogram of visited energies is increased by $H(E_B) \to H(E_B) + 1$. If the trial move was rejected, $g'(E_A)$ and $H(E_A)$ are updated instead in the same way. Once $H(E)$ is "flat," the modification factor f will be reduced and all histogram entries will be reset to zero. The method performs this procedure iteratively until f is less than some predefined threshold value f_{final}. Even though multiple improvements of the details of this sampling method exist, we stick to the original procedure, where the initial modification factor is set to $f_{\mathrm{init}} = e^1$ and decreased via $\ln f \to \ln f/2$, and $\ln f_{\mathrm{final}} = 1 \times 10^{-8}$. The initial guess for $g'(E)$ is $g'(E) = 1$ and we use the "80%" flatness criterion for the histogram $H(E)$. That is, all $H(E)$ entries are no less than 80% of the mean histogram height. Eventually, if we avoid potential systematic errors, $g'(E)$ will converge to the true $g(E)$ [53].

The trial moves we adopted in our simulation are pull moves [54] and bond-rebridging moves [55]. It has been found that these two trial moves work amazingly well with Wang-Landau sampling for both the determination of the minimum energy state and the estimation of the density of states for the model used here [40,41]. In our simulation, we used move fractions of 75% and 25% for pull moves and bond-rebridging moves, respectively.

C. Thermodynamic and structural quantities

The partition function $Z(T)$ of the system at a particular temperature T can be obtained with the knowledge of the density of states, $g(E)$:

$$Z(T) = \sum_E g(E)\, e^{-E/k_{\mathrm{B}}T}, \tag{3}$$

where k_{B} is the Boltzmann constant. This allows us to calculate the thermal properties such as the mean energy $\langle E\rangle(T)$, the heat capacity $C_V(T)$, and the ground-state population $P_0(T)$ [44]:

$$C_V(T) = \frac{\langle E^2\rangle - \langle E\rangle^2}{k_{\mathrm{B}}T^2}, \tag{4}$$

$$P_0(T) = \frac{g(E_0)\, e^{-E_0/k_{\mathrm{B}}T}}{Z(T)}, \tag{5}$$

where E_0 is the ground-state energy. As is common when studying generic models, we work in reduced units in the following; i.e., we set $k_{\mathrm{B}} = 1$ and $\epsilon_{\mathrm{HH}} = 1$.

Besides those thermodynamic quantities, structural observables are also important in understanding the conformational changes during the folding process. We measure two commonly used quantities, radius of gyration (R_{g}) and end-to-end distance (R_{ee}):

$$R_{\mathrm{g}} = \left(\frac{1}{N}\sum_{i=1}^{N}(\vec{r}_i - \vec{r}_{\mathrm{c.m.}})^2\right)^{1/2}, \tag{6}$$

$$R_{\mathrm{ee}} = |\vec{r}_N - \vec{r}_1|, \tag{7}$$

where N is the number of monomers in the chain; $\vec{r}_i$ and $\vec{r}_{\mathrm{c.m.}}$ represent the positions of the ith monomer and the center of mass of the given configuration, respectively.

In addition, Wüst and Landau [41] proposed another scalar structural observable, the tortuosity τ of the protein:

$$\tau = \left(\frac{1}{N-2}\sum_{i=1}^{N-2}(s_i - \bar{s})^2\right)^{1/2}, \tag{8}$$

where

$$s_i = \sum_{j=1}^{i} \vec{r}_{j,j+1} \times \vec{r}_{j,j+2}, \quad 1 \leqslant i \leqslant N-2. \tag{9}$$

Here $\vec{r}_{j,j+1}$ (or $\vec{r}_{j,j+2}$) denotes a vector pointing from monomer j to $j+1$ (or $j+2$); $\bar{s}$ is the average of s_i. Unlike the radius of gyration and the end-to-end distance, which measure spatial extent only, τ is particularly sensitive to sequence-dependent internal topological features such as the breaking of HH contacts in compact denatured states upon folding to the ground state. For our purpose, it serves as a complementary structural quantity to better interpret features in the specific heat curves, for example. See also Ref. [41] for a discussion of this observable in the context of lattice polymers. To generally obtain more accurate structural quantities, we adopt the Wang-Landau resampling procedure, also proposed in Ref. [41].

D. Characterization of ground-state structures

One of the difficulties included in studying the ground-state properties of proteins in the HP model (especially for long sequences) is the enormous degeneracy and high symmetry on the simple cubic lattice. Here we devise a simple heuristic method to characterize and store ground-state structures during the simulation.

1. *Sequence of directions*

For a given HP protein conformation on the simple cubic lattice, the "path" from the first monomer through the end can be uniquely recorded as a sequence of directions. We define two sets of values, $\vec{B}_1,\vec{B}_2,\ldots,\vec{B}_{N-1}$ and $D_1,D_2,\ldots,D_{N-1}$, for the $N-1$ bonds that connect consecutive monomers in a sequence of length N. The former one, for instance $\vec{B}_k$, is determined by the difference of coordinates between monomers k and $k+1$. Therefore, $\vec{B}_k$ will be assigned one of the values from $\{+\vec{X},-\vec{X},+\vec{Y},-\vec{Y},+\vec{Z},-\vec{Z}\}$, where $+\vec{X}$ denotes the positive X-axis direction, etc. The latter one is the sequence of direction (SoD), which contains five elements: *forward* (F), *left* (L), *right* (R), *up* (U), and *down*(D) (see Refs. [56,57] for similar representations). The procedure of calculating the sequence of directions can be described as follows:

(1) Along the HP chain, pick the first three bonds $(1,i,j)$ which are all perpendicular to each other, i.e., such that $\vec{B}_1 \perp \vec{B}_i \perp \vec{B}_j \perp \vec{B}_1$.

(2) Then $\vec{B}_1$, $\vec{B}_i$, and $\vec{B}_j$ define a new coordinate system (i.e., new directions $+\vec{X}$, $+\vec{Y}$, and $+\vec{Z}$) in which we calculate the remaining bonds.

(3) The first bond (D_1) is, by definition, the *forward* direction, while the next nonforward bond (D_i) is defined as *left*. The other directions are then determined in step 4.

(4) Assign F to $D_2,\ldots,D_{i-1}$ and calculate D_h, $i<h<N$:

$$D_h = \begin{cases} F, & \text{IF} \quad \vec{B}_h = \vec{B}_{h-1} \\ L, & \text{ELIF} \quad \vec{B}_h = \vec{B}_{h-k} \\ & \text{AND} \quad o(\vec{B}_{h-1},\vec{B}_h) = s(\vec{B}_{h-1},\vec{B}_h) \\ R, & \text{ELIF} \quad \vec{B}_h = \vec{B}_{h-k} \\ & \text{AND} \quad o(\vec{B}_{h-1},\vec{B}_h) \neq s(\vec{B}_{h-1},\vec{B}_h) \\ U, & \text{ELIF} \quad o(\vec{B}_{h-k},\vec{B}_{h-1}) = s(\vec{B}_{h-1},\vec{B}_h) \\ & \text{AND} \quad s(\vec{B}_h,+) = 1 \\ U, & \text{ELIF} \quad o(\vec{B}_{h-k},\vec{B}_{h-1}) \neq s(\vec{B}_{h-1},\vec{B}_h) \\ & \text{AND} \quad s(\vec{B}_h,-) = 1 \\ D, & \text{ELIF} \quad o(\vec{B}_{h-k},\vec{B}_{h-1}) = s(\vec{B}_{h-1},\vec{B}_h) \\ & \text{AND} \quad s(\vec{B}_h,-) = 1 \\ D, & \text{ELIF} \quad o(\vec{B}_{h-k},\vec{B}_{h-1}) \neq s(\vec{B}_{h-1},\vec{B}_h) \\ & \text{AND} \quad s(\vec{B}_h,+) = 1, \end{cases}$$

where $\vec{B}_{h-k}$ is the closest preceding element that satisfies $\vec{B}_{h-k} \neq \vec{B}_{h-1}$ and $o(\vec{B}_m,\vec{B}_n)$, $s(\vec{B}_m,\vec{B}_n)$ are functions defined as

$$o(\vec{B}_m,\vec{B}_n) = \begin{cases} 1\,, & (|\vec{B}_m|,|\vec{B}_n|) \in \{(X,Y),(Y,Z),(Z,X)\} \\ 0\,, & \text{otherwise,} \end{cases}$$

$$s(\vec{B}_m,\vec{B}_n) = \begin{cases} 1\,, & \vec{B}_m,\ \vec{B}_n \text{ have the same sign} \\ 0, & \text{otherwise.} \end{cases}$$

By this procedure we uniquely assign a SoD to each conformation and vice versa, taking the symmetries into account. That is, conformations are equal (modulo symmetry transformation of the cubic lattice) if and only if their SoD are identical.

2. *Ground-state sampling*

To obtain all the ground-state structures of a given HP sequence we perform a multicanonical sampling [58] on the whole energy space. During that process, trial states are generated as before and also accepted or rejected according to Eq. (2), where $g'(E)$ is now the final estimator obtained from the preceding Wang-Landau run and not updated anymore. If the putative ground-state energy, i.e., the lowest energy found during the Wang-Landau run, is met, we calculate the direction sequence of this state and compare it to those of previously found ground-state structures, which we store in a tree structure container with a branching factor of at most 5 (the number of elements in a SoD). In this tree data structure, a direction sequence is uniquely represented by a path of length $N-1$ from the root node to a leaf node. Hence, the complexity of verifying a new found direction sequence is $O(N)$ [59]. If the actual ground state is already present in that container, we just proceed. Otherwise, the new structure will be added to the database and the counter of degeneracy increased. The simulation ideally ends when the rate of finding new ground states approaches zero; i.e, the estimator for the ground-state degeneracy converges. In practice, we terminate the runs after a predefined number of MC steps (see below). Note that even though we use this method mainly to estimate ground-state degeneracies, it is of course applicable to any other energy level just as well.

III. RESULTS AND DISCUSSION

A. Ground-state structure searching

As a verification of our method, we chose four prominent HP sequences with length of 14, all of which have been studied using an enumeration method [21], and performed a multicanonical scan counting the absolute density of states. That is, we estimate the degeneracy of all energy levels analogously to the ground-state sampling described above. The results are shown in Table I, where the numbers of unique structures at each energy level are given. By identifying the dimension of each structure and considering different symmetries (1D × 6; 2D × 24; 3D × 48), we calculated the densities of states and found them to be exactly the same compared to enumeration results [21]. Each of these simulations took fewer than 4×10^9 Monte Carlo steps for $g(E_0)$ to converge.

After this proof of concept, some other widely studied HP sequences [56,60,61] have been chosen for testing our scheme. We carried out simulations for estimating the ground-state

TABLE I. Our Monte Carlo results for absolute densities of states for four 14mers. Columns from left to right: energy level, total number of structures of all dimensions, 2D structures (n_{2D}), 3D structures (n_{3D}) and $g(E) = 6\,n_{1D} + 24\,n_{2D} + 48\,n_{3D}$. Each sequence has only 1 1D-structure (with $E = 0$) which is not shown. Our results are identical to results from exact enumeration [21].

SeqID: 14.1 (HPHPHHHPHPHHPPH)

E	All	n_{2D}	n_{3D}	$g(E)$
−8	1	0	1	48
−7	262	0	262	12 576
−6	3 380	5	3 375	162 120
−5	28 163	84	28 079	1 349 808
−4	176 076	713	175 363	8 434 536
−3	754 422	3 809	7 50 613	36 120 840
−2	2 466 457	14 059	2 452 398	118 052 520
−1	6 533 719	38 605	6 495 114	312 691 992
0	9 758 750	52 912	9 705 837	467 150 070
SUM	19 721 230			943 974 510

SeqID: 14.2 (HHPPHPHPHHPHPH)

E	All	n_{2D}	n_{3D}	$g(E)$
−8	2	0	2	96
−7	220	0	220	10 560
−6	2 929	4	2 925	140 496
−5	22 738	68	22 670	1 089 792
−4	139 052	561	138 491	6 661 032
−3	625 336	3 014	622 322	29 943 792
−2	2 102 592	10 872	2 091 720	100 663 488
−1	5 710 617	31 935	5 678 682	273 343 176
0	11 117 744	63 733	11 054 010	532 122 078
SUM	19 721 230			943 974 510

SeqID: 14.3 (HHPHPHPPHPHPHH)

E	All	n_{2D}	n_{3D}	$g(E)$
−8	2	0	2	96
−7	200	1	199	9 576
−6	2 631	2	2 629	126 240
−5	21 987	68	21 919	1 053 744
−4	125 858	510	125 348	6 028 944
−3	591 753	3 110	588 643	28 329 504
−2	2 286 507	13 296	2 273 211	109 433 232
−1	6 392 045	37 796	6 354 249	305 911 056
0	10 300 247	55 404	10 244 842	493 082 118
SUM	19 721 230			943 974 510

SeqID: 14.4 (HHPHPPPHPHPHHPH)

E	All	n_{2D}	n_{3D}	$g(E)$
−8	4	0	4	192
−7	232	0	232	11 136
−6	3 348	7	3 341	160 536
−5	26 267	74	26 193	1 259 040
−4	163 540	757	162 783	7 831 752
−3	801 505	4 370	797 135	38 367 360
−2	2 702 687	15 734	2 686 953	129 351 360
−1	6 575 905	39 087	6 536 818	314 705 352
0	9 447 742	50 158	9 397 583	452 287 782
SUM	19 721 230			943 974 510

degeneracy for each of these sequences and listed the results in Table II. For short sequences (e.g., 27.2, 27.3, and 31) or sequences with low ground-state degeneracy (e.g., 42 and 67), the results of our simulation agree with other studies perfectly.

TABLE II. Estimated ground-state degeneracy of some widely studied HP proteins. For each of them we listed the ground-state energy E_0, the ground-state degeneracy $g^{\mathrm{L}}(E_0)$ found in earlier studies, and $g(E_0)$ estimated with our method. Converged sequences do not have statistical errors; otherwise error bars were obtained from multiple extrapolation fits (see text).

SeqID	E_0	$g^{\mathrm{L}}(E_0)$	$g(E_0)$
27.1	−16	36691[a]	51537
27.2	−15	297[a]	297
27.3	−16	25554[a]	25554
31	−28	1114[a]	1114
42	−34	4[b]	4
67	−56	3[b]	3
48.1	−32	$(5.2 \pm 0.8)\times 10^{6}$[c]	$(10.3 \pm 0.4) \times 10^{6}$
48.2	−34	$(1.7 \pm 0.8) \times 10^{4}$[c]	$(2.84 \pm 0.02)\times 10^{4}$
48.3	−34	$(6.6 \pm 2.8) \times 10^{3}$[c]	5.09×10^{3}
48.4	−33	$(6.0 \pm 1.3) \times 10^{4}$[c]	$(4.97 \pm 0.16)\times 10^{4}$
48.5	−32	$(1.2 \pm 0.3) \times 10^{6}$[c]	$(1.94 \pm 0.04)\times 10^{6}$
48.6	−32	$(9.6 \pm 1.9) \times 10^{4}$[c]	$(1.84 \pm 0.02)\times 10^{6}$
48.7	−32	$(5.8 \pm 2.1) \times 10^{4}$[c]	$(10.8 \pm 0.1)\times 10^{4}$
48.8	−31	$(2.2 \pm 0.7) \times 10^{7}$[c]	$(1.59 \pm 0.03)\times 10^{7}$
48.9	−34	$(1.4 \pm 0.5) \times 10^{3}$[c]	2.614×10^{3}
48.10	−33	$(1.9 \pm 0.9) \times 10^{5}$[c]	$(5.53 \pm 0.14)\times 10^{5}$

[a] Values of $g^{\mathrm{L}}(E_0)$ taken from Yue and Dill [56].
[b] Values of $g^{\mathrm{L}}(E_0)$ taken from Yue and Dill [60].
[c] Values of $g^{\mathrm{L}}(E_0)$ taken from Bachmann and Janke [61].

Due to the high ground-state degeneracy, it is extremely challenging to reach all ground states for the 48mers [61] in finite simulation time (we ran up to 2.5 × 10^{11} MC steps). However, there are two sequences (48.3 and 48.9) for which the ground-state degeneracy stayed stable for a long time, and we thus believe we have converged to the true value $g(E_0)$. By normalizing the number of ground states and Monte Carlo time, we find that these two sequences share the same convergence behavior. The assumption of a fundamental convergence pattern provides a means to extrapolate the true ground-state degeneracy for other long sequences. Hence, we fitted the normalized number of visited, different ground states vs time for other sequences to the known curve for 48.3 and extrapolated their ground-state degeneracy. As an example, we show in Fig. 2 the corresponding fit for sequence 48.1. The extrapolated part is marked by the dashed line in the figure. However, instead of a unique fit, there are multiple choices which fit equally well. Therefore, by doing 20 different fits, the average value as well as an error bar could be calculated. Even though not every curve fits as well as Fig. 2, it provides a better estimation of the true value $g(E_0)$. We note that our procedure yields rather different values than those obtained earlier using an approach where $g(E_0)$ is obtained from an implicit estimate of the partition function [61]. However, since our estimates are obtained from explicit enumeration of ground states with very high statistics, we believe that our procedure provides more reliable results.

B. Effect of single-site mutations

After this instructive preparatory work, we focus on the main part of this study: the effect of single-site mutations on

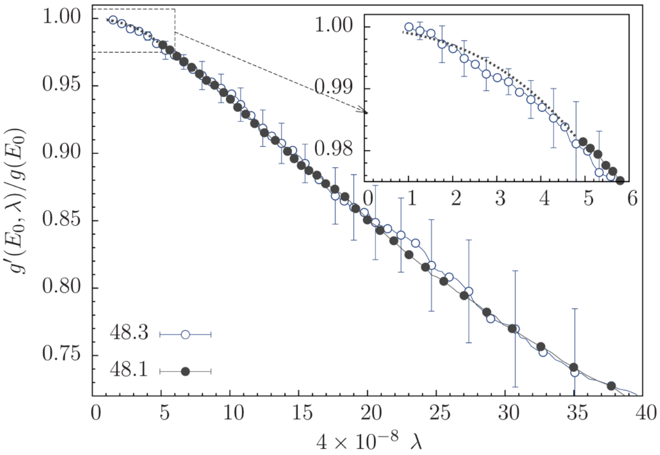

FIG. 2. (Color online) Number of different ground states found over time for HP sequences 48.1 fitted to the corresponding curve of protein 48.3. $g'(E_0,\lambda)$ is the actual estimator of ground-state degeneracy and $\lambda = 1/(10\,000$ MC steps) is the inverse Monte Carlo time. For clarity, only selected data points and error bars are shown. See text for details.

ground states and the thermodynamic behavior of HP proteins. A single-site mutation (SSM) on a HP protein consists of a “flip” of one monomer from its original type to the other. For example, HPHHP becomes HPPHP under SSM on the third monomer. To understand possible effects of SSM on HP proteins, we choose two sequences which were designed to study the origins of tertiary structures in proteins [60]:

$$\text{Seq3D42} : \mathrm{P(H(HP)_2)_2HP_2H_3(PH)_2(HP)_2H_3P_2H((PH)_2H)_2P},$$

$$\text{Seq3D67} : \mathrm{(PH(PH_2)_2PHP_2H_3PP)_3PH(PH_2)_2PHP_2H_3P}$$

$$\equiv x\mathrm{P}x\mathrm{P}x\mathrm{P}x, \quad \text{where } x = \mathrm{PH(PH_2)_2PHP_2H_3P}.$$

There are symmetries present in these two sequences: Seq3D42 reads the same forward and backward, and Seq3D67 is composed of four identical pieces, each pair of which is connected through a P monomer. The ground-state structures of these two lattice proteins mimic by construction α/β barrels and the β helix, respectively (see Appendix). Note that the ground-state degeneracies are extremely small (Table II). The SSMs have been systematically performed on each monomer of both HP chains. We thus create 42 and 67 mutated sequences, respectively. We denote Seq3D42s*k* as the mutated sequence generated by applying a SSM on the *k*th monomer of Seq3D42 (and analogously for Seq3D67). We performed simulations of each of these mutated sequences independently as described earlier. Results are shown and discussed in the following.

1. Ground states

We are first interested in how SSMs affect the ground states (GSs) of HP proteins in terms of their energies, actual conformations, and degeneracies. In Figs. 3 and 4 we plot the ground-state energies and degeneracies for all SSMs of Seq3D42 and Seq3D67, respectively. For both sequences we find that the effect of SSMs can vary significantly, depending on the monomer position. About half of the mutated sequences retain their ground-state energy (GSE), while others changed significantly. For example, mutations on the 15th monomer of Seq3D42 or the 13th of Seq3D67 change the GSEs from $E = -34$ to -30 and from $E = -56$ to -51, respectively. Moreover, three mutated sequences (Seq3D67s17, Seq3D67s34, and Seq3D67s51) even have a lower GSE ($E = -57$) compared to the original sequence. Interestingly, none of the mutated sequences has a GSE of $E = -33$ or -55, which correspond to the first excited states of the unmutated Seq3D42 and Seq3D67, respectively.

Regarding the ground-state degeneracy (GSD), most of the mutated sequences show dramatically larger values than the unmutated ones. However, by comparing their ground-state structures, we found that 88 out of 109 (36 for the 42mer and 52 for the 67mer) mutated sequences retain the ground-state structures of the original sequence. For Seq3D67 we identified six P monomers (at positions 11, 17, 28, 34, 45, and 51; cf. Fig. 4) which are “immune” against SSM in the sense that the ground-state degeneracy and the actual ground-state structures stay exactly the same except for the substituted site. For three of them (17, 34, and 51), SSM even results in lower ground-state

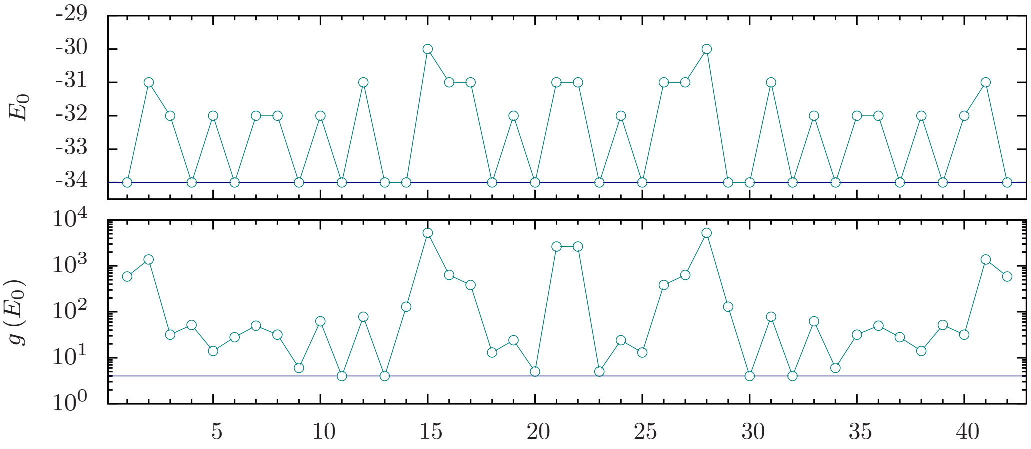

FIG. 3. (Color online) Ground-state energy E_0 (top) and ground-state degeneracy $g(E_0)$ (bottom) of mutated Seq3D42. The X-axis value indicates the position which has been affected by the single-site mutation. Properties of the original, unmutated sequence are marked by horizontal lines.

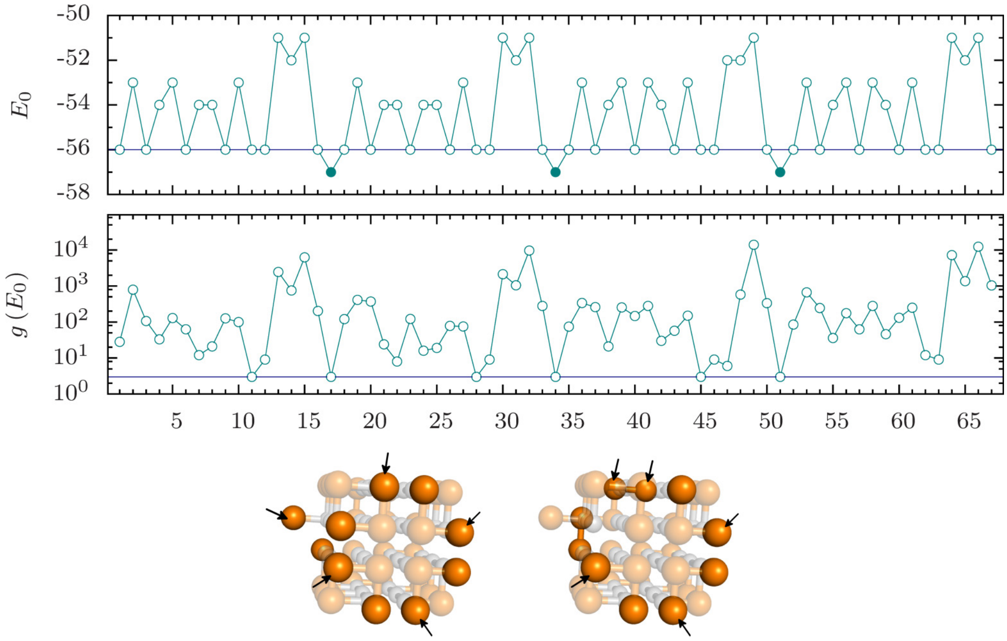

FIG. 4. (Color online) Ground-state energy E_0 and ground-state degeneracy $g(E_0)$ of mutated Seq3D67 (cf. Fig. 3). Each of the bottom pictures shows the result of two overlapping ground-state structures of Seq3D67. Overlapped monomers are shown in faint color. Monomers pointed to by arrows belong to the same structure, while the rest belong to different structures.

energies. Furthermore, we saw that the ground states remain unaffected under multiple site mutations, i.e., mutating up to all three sites simultaneously. We also find sequences where a single-site mutation affects not the ground-state energy but its degeneracy. In these cases we observe that SSM lowers the thermal stability of ground states by notably increasing the degeneracy of the first excited states, for example. Through examining the 3D ground-state structures of the 67mer, we found that these "immune" positions are at the joints of lattice helices and lattice strands, while those extremely sensitive sites (e.g., 15, 32, and 49) are usually located at the lattice strands (cf. Appendix). We note that the symmetries observed in Fig. 3 reflect the symmetry in the HP sequence of Seq3D42 as expected, providing further evidence for the validity of our method.

In Fig. 5, we plot the specific heat $C_V(T)/N$ and the ground-state population $P_0(T)$ [Eqs. (4) and (5)] of two sequences for which both the ground-state energy and the ground-state degeneracy do not change compared to the corresponding unmutated sequence. We see a shift of the ground-state population P_0 to lower temperatures. For example, at the temperatures where 50% of the conformations in the canonical distributions of the unmutated sequences correspond to ground states, this percentage drops to 17–18% for both mutated sequences (see grid lines in Fig. 5). That is, the ground-state population is much more sensitive to temperature increase compared to the original protein. Looking at the heat capacity, we also note a shift of the low-energy peak which corresponds to the formation of a compact hydrophobic core; i.e., coming from the ground state, this core breaks apart at lower temperatures as an effect of the mutation. Both observations show that the mutations lower the thermal stability of the ground states of the investigated HP proteins.

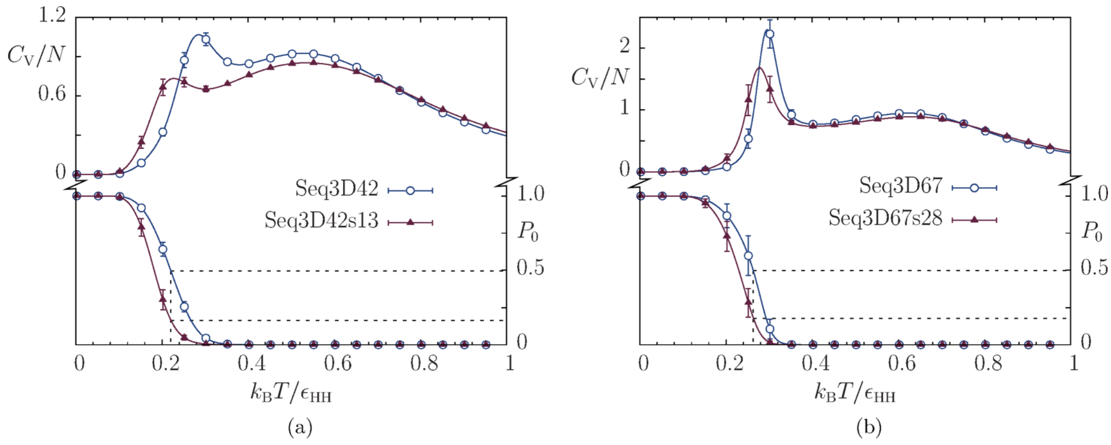

FIG. 5. (Color online) Examples of thermal stability of the ground state: (a) comparison between Seq3D42s13 and Seq3D42 based on specific heat (top curves, left-hand scale) and ground-state population (bottom curves, right-hand scale) and (b) comparison between Seq3D67s28 and Seq3D67 based on specific heat and ground-state population. In both figures, error bars smaller than data points are not shown.

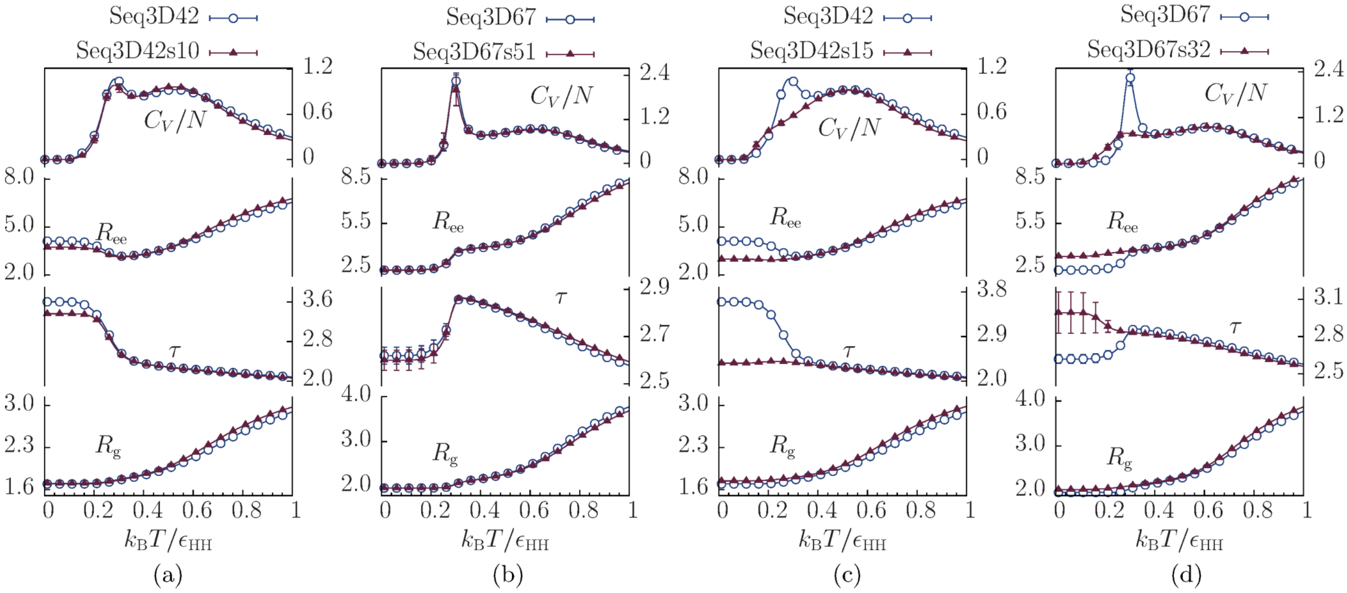

FIG. 6. (Color online) Effect of mutations on folding behavior: (a, b) two cases where the mutations do not affect the folding behavior and (c, d) examples of changed thermodynamic quantities under mutation. In all panels above, error bars smaller than data points are not shown.

2. *Thermodynamic and structural properties*

Finally, we investigate in more detail how single-site mutations can affect the thermal behavior of HP proteins. Such knowledge could help unveil the effect of mutations on the folding process, for example. The quantities we are interested in here are the specific heat, the end-to-end distance, the tortuosity, and the radius of gyration, as defined in Sec. II C.

By examining all mutated sequences of Seq3D42 and (most of) Seq3D67, we have discovered cases where all quantities revealed very similar behavior compared to the unmutated sequences [see Figs. 6(a) and 6(b)]. This comparison strongly indicates that those mutations do not affect the folding behavior significantly. More than 50% of all single-site mutations fall into this class, in which sequences contain the original ground-state structures and add, if at all, only a small number to the ground-state degeneracy. On the other hand, there are instances where mutations significantly affect thermal quantities. As shown in Figs. 6(c) and 6(d), mutated sequences can present different behaviors in various ways. Typically, the left-hand peak of heat capacity, which corresponds to the hydrophobic core formation (see, for example, Ref. [61] for a more detailed discussion of this transition), becomes lower or fades into a shoulder, along with raised or lowered τ and R_{ee}. Significant change of R_{g} has not been observed in all of our cases, which implies that this quantity is, not surprisingly for this model, quite stable under single-site mutation. Under this type of effects, mutated sequences also show a sharp increase in their ground-state degeneracies and might lose the original ground-state structures.

IV. SUMMARY AND CONCLUSION

The effect of mutations on proteins is of fundamental interest in many areas of life sciences. There are different approaches to study this effect by means of computer simulations, leading to complementary insights: one could choose an atomistic model for a specific protein and study a specific mutation, or one could choose generic models and perform systematic studies of general mechanisms. There have been a number of works for the latter approach [22,42–44], but we provide a distinct method to systematically and thoroughly study mutations on large lattice proteins, i.e., proteins that are larger than ∼30 monomers and which will not be accessible by exact enumeration any time soon, in a conclusive and reliable manner.

Our heuristic method estimates the density of states, including the absolute ground-state degeneracy, of HP lattice proteins. It combines flat-histogram sampling with an efficient structure database and enables us to gain detailed insight into systems of sizes far beyond those accessible by enumeration approaches, while also working as effectively as such methods for short HP sequences. Moreover, Wang-Landau sampling with appropriate trial moves has proven to be successful in ground-state searching for HP sequences as long as 136 monomers [41]. Therefore, we believe our method should be at least suitable for sequences with this length. To demonstrate the usefulness of this method, we applied it to thoroughly investigate the effect of single-site mutation on two long, designed HP proteins and discovered that many mutations do not affect the protein significantly in any regard, including the ground-state degeneracy and energy. On the other hand, very sensitive positions in the primary structure exist, where mutations can drastically change the folding process and low-energy structures. Remarkably, both observations coincide with experimental discoveries for real proteins [5,7], as discussed in Sec. I, confirming the adequacy of simple, generic models for certain problems. In addition, we found that the thermal stability of mutated sequences is likely to be lower than that for original sequences, from the observation of ground-state population and specific heat. The reason is that even though the ground-state degeneracy may increase dramatically after mutations, the degeneracies of the first and second excited states grow with the same rate or even faster.

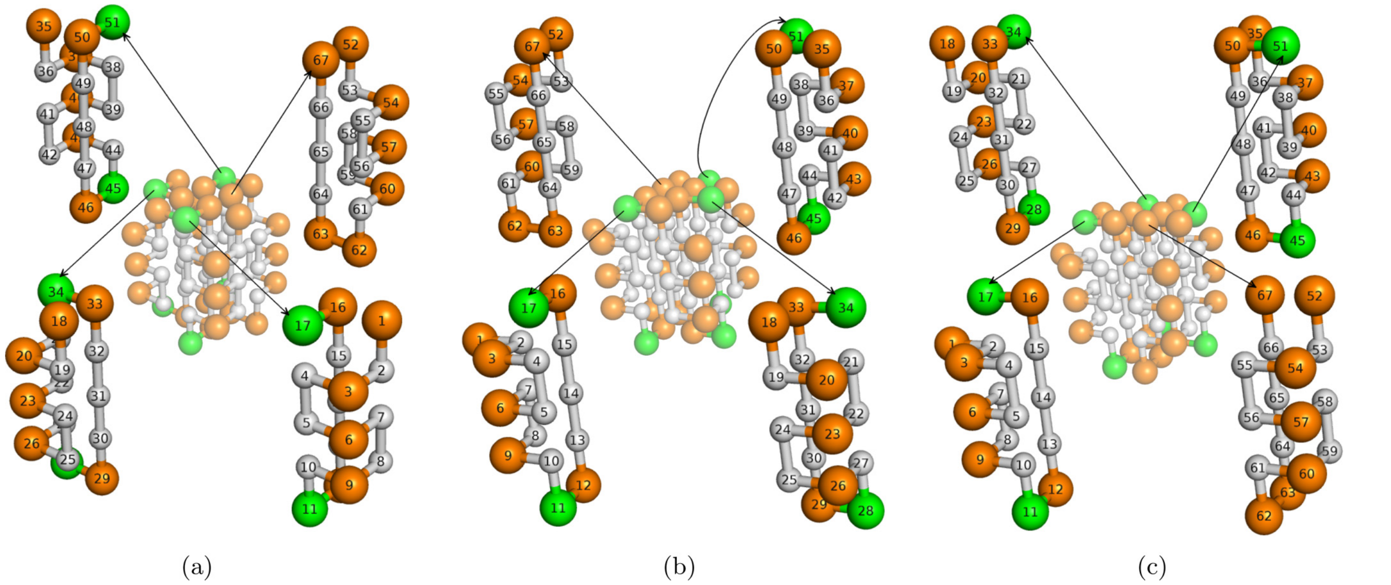

FIG. 7. (Color online) Three different ground-state structures for the 67mer studied in this paper. Residues 2–10 form a lattice helix, and residues 12–16 form a lattice strand [60]. Hydrophobic monomers are represented by small, silver beads, while polar monomers are represented by big beads which are either orange (light gray) or green (dark). Green monomers are those for which the ground-state degeneracy remains the same under single-site mutations; they are located at the joints connecting lattice strands and lattice helices.

ACKNOWLEDGMENTS

We thank M. Bachmann for constructive discussions. This work was supported by the National Science Foundation under Grants No. DMR-0810223 and No. OCI-0904685. Y.W.L. was sponsored by the Office of Advanced Scientific Computing Research, U.S. Department of Energy Contract No. DE-AC05-00OR22725. Part of the work was carried out at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725. Computing resources were partially provided by the Georgia Advanced Computing Resource Center and the Texas Advanced Computing Center under XSEDE Grant No. PHY130014. LA-UR-14-21619 assigned.

APPENDIX: GROUND-STATE STRUCTURES

Three different ground-state structures are shown for the 67mer (Fig. 7) and four different ground-state structures for the 42mer (Fig. 8) studied in this paper.

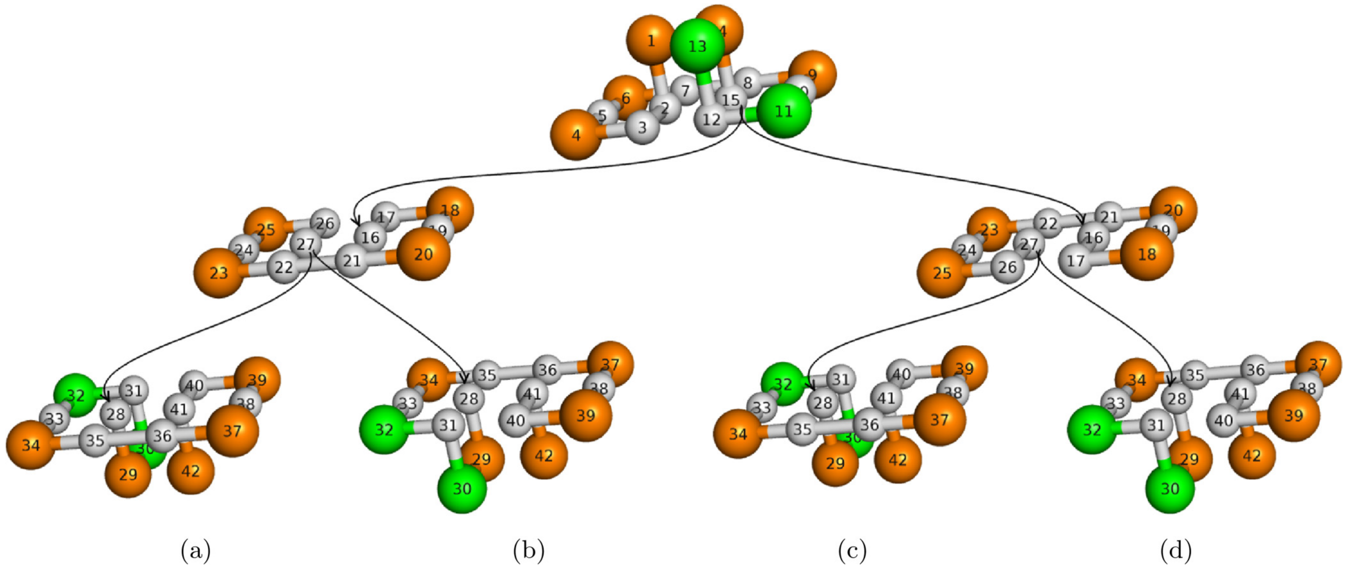

FIG. 8. (Color online) Four different ground-state structures for the 42mer studied in this paper. Each ground-state structure is sliced into three layers. The top layer is shared by all the ground-state structures, while there are two different structures for the middle and bottom layers, respectively. Arrows in the figure point to the bonded monomer in the next layer. Hydrophobic monomers are represented by small, silver beads, while polar monomers are represented by big beads which are either orange (light gray) or green (dark). Green colored monomers are those for which the ground-state degeneracy remains the same under single-site mutation.

[1] K. A. Dill and J. L. MacCallum, Science **338**, 1042 (2012).
[2] A. R. Davidson, Proc. Natl. Acad. Sci. USA **105**, 2759 (2008).
[3] C. G. Roessler, B. M. Hall, W. J. Anderson, W. M. Ingram, S. A. Roberts, W. R. Montfort, and M. H. J. Cordes, Proc. Natl. Acad. Sci. USA **105**, 2343 (2008).
[4] P. A. Alexander, Y. He, Y. Chen, J. Orban, and P. N. Bryan, Proc. Natl. Acad. Sci. USA **104**, 11963 (2007).
[5] S. Rackovsky, Phys. Rev. Lett. **106**, 248101 (2011).
[6] J. H. Miller, C. Coulondre, M. Hofer, U. Schmeissner, H. Sommer, A. Schmitz, and P. Lu, J. Mol. Biol. **131**, 191 (1979).
[7] J. U. Bowie, J. F. Reidhaar-Olson, W. A. Lim, and R. T. Sauer, Science **247**, 1306 (1990).
[8] K. A. Dill, Protein Sci. **8**, 1166 (1999).
[9] A. Kolinski and J. Skolnick, Polymer **45**, 511 (2004).
[10] K. A. Dill, Biochemistry **24**, 1501 (1985).
[11] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[12] V. Castells, S. Yang, and P. R. Van Tassel, Phys. Rev. E **65**, 031912 (2002).
[13] M. Bachmann and W. Janke, Phys. Rev. E **73**, 020901 (2006).
[14] A. Swetnam and M. P. Allen, Phys. Rev. E **85**, 062901 (2012).
[15] M. Radhakrishna, S. Sharma, and S. K. Kumar, J. Chem. Phys. **136**, 114114 (2012).
[16] Y. W. Li, T. Wüst, and D. P. Landau, Phys. Rev. E **87**, 012706 (2013).
[17] R. Bonaccini and F. Seno, Phys. Rev. E **60**, 7290 (1999).
[18] G. Ping, J. M. Yuan, M. Vallieres, H. Dong, Z. Sun, Y. Wei, F. Y. Li, and S. H. Lin, J. Chem. Phys. **118**, 8042 (2003).
[19] B. Pattanasiri, Y. W. Li, D. P. Landau, T. Wüst, and W. Triampo, J. Phys.: Conf. Ser. **402**, 012048 (2012).
[20] A. Irbäck and C. Troein, J. Biol. Phys. **28**, 1 (2002).
[21] M. Bachmann and W. Janke, Acta Phys. Pol. B **34**, 4689 (2003).
[22] R. Schiemann, M. Bachmann, and W. Janke, J. Chem. Phys. **122**, 114705 (2005).
[23] S. L. Narasimhan, A. K. Rajarajan, and L. Vardharaj, J. Chem. Phys. **137**, 115102 (2012).
[24] R. Backofen and S. Will, Constraints **11**, 5 (2006).
[25] E. M. O'Toole and A. Z. Panagiotopoulos, J. Chem. Phys. **97**, 8644 (1992).
[26] T. C. Beutler and K. A. Dill, Protein Sci. **5**, 2037 (1996).
[27] P. Grassberger, Phys. Rev. E **56**, 3682 (1997).
[28] H. Frauenkron, U. Bastolla, E. Gerstner, P. Grassberger, and W. Nadler, Phys. Rev. Lett. **80**, 3149 (1998).
[29] M. Bachmann and W. Janke, Phys. Rev. Lett. **91**, 208105 (2003).
[30] H.-P. Hsu, V. Mehra, W. Nadler, and P. Grassberger, J. Chem. Phys. **118**, 444 (2003); Phys. Rev. E **68**, 021113 (2003).
[31] T. Prellberg and J. Krawczyk, Phys. Rev. Lett. **92**, 120602 (2004).
[32] J. L. Zhang and J. S. Liu, J. Chem. Phys. **117**, 3492 (2002).
[33] J. Zhang, S. C. Kou, and J. S. Liu, J. Chem. Phys. **126**, 225101 (2007).
[34] W. Tang and Q. Zhou, Phys. Rev. E **86**, 031909 (2012).
[35] R. Unger and J. Moult, J. Mol. Biol. **231**, 75 (1993).
[36] R. König and T. Dandekar, BioSystems **50**, 17 (1999).
[37] F. Liang and W. H. Wong, J. Chem. Phys. **115**, 3374 (2001).
[38] A. Shmygelska and H. H. Hoos, BMC Bioinf. **6**, 30 (2005).
[39] F. Wang and D. P. Landau, Phys. Rev. Lett. **86**, 2050 (2001); Phys. Rev. E **64**, 056101 (2001); D. P. Landau, S.-H. Tsai, and M. Exler, Am. J. Phys. **72**, 1294 (2004).
[40] T. Wüst and D. P. Landau, Phys. Rev. Lett. **102**, 178101 (2009).
[41] T. Wüst and D. P. Landau, J. Chem. Phys. **137**, 064903 (2012).
[42] K. F. Lau and K. A. Dill, Proc. Natl. Acad. Sci. USA **87**, 638 (1990).
[43] D. Shortle, H. S. Chan, and K. A. Dill, Protein Sci. **1**, 201 (2008).
[44] C. Holzgräfe, A. Irbäck, and C. Troein, J. Chem. Phys. **135**, 195101 (2011).
[45] F. J. Blanco, I. Angrand, and L. Serrano, J. Mol. Biol. **285**, 741 (1999).
[46] S. Dalal and L. Regan, Protein Sci. **9**, 1651 (2000).
[47] T. A. Anderson, M. H. J. Cordes, and R. T. Sauer, Proc. Natl. Acad. Sci. USA **102**, 18344 (2005).
[48] X. I. Ambroggio and B. Kuhlman, Curr. Opin. Struct. Biol. **16**, 525 (2006).
[49] Y. He, Y. Chen, P. Alexander, P. N. Bryan, and J. Orban, Proc. Natl. Acad. Sci. USA **105**, 14412 (2008).
[50] P. A. Alexander, Y. He, Y. Chen, J. Orban, and P. N. Bryan, Proc. Natl. Acad. Sci. USA **106**, 21149 (2009).
[51] W. Kauzmann, in *Advances in Protein Chemistry*, Vol. 14, edited by C. B. Anfinsen, M. L. Anson, K. Bailey, and J. T. Edsall (Academic Press, New York, 1959), p. 1.
[52] K. A. Dill, Science **250**, 297 (1990).
[53] C. Zhou and R. N. Bhatt, Phys. Rev. E **72**, 025701 (2005).
[54] N. Lesh, M. Mitzenmacher, and S. Whitesides, in *RECOMB: Proceedings of the Seventh Annual International Conference on Computational Biology, Berlin, Germany* (Association for Computing Machinery, New York, 2003), p. 188.
[55] J. M. Deutsch, J. Chem. Phys. **106**, 8849 (1997).
[56] K. Yue and K. A. Dill, Phys. Rev. E **48**, 2267 (1993).
[57] R. Schiemann, M. Bachmann, and W. Janke, Comput. Phys. Commun. **166**, 8 (2005).
[58] B. A. Berg and T. Neuhaus, Phys. Lett. B **267**, 249 (1991); Phys. Rev. Lett. **68**, 9 (1992).
[59] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. (MIT Press, Cambridge, MA, 2009).
[60] K. Yue and K. A. Dill, Proc. Natl. Acad. Sci. USA **92**, 146 (1995).
[61] M. Bachmann and W. Janke, J. Chem. Phys. **120**, 6779 (2004).