A Study of Density of States and Ground States in Hydrophobic-Hydrophilic Protein Folding Models by Equi-energy Sampling

A study of density of states and ground states in hydrophobic-hydrophilic protein folding models by equi-energy sampling

S. C. Kou[a] and Jason Oh
Department of Statistics, Science Center, Harvard University, Cambridge, Massachusetts 02138

Wing Hung Wong
Department of Statistics, Sequoia Hall, Stanford University, Stanford, California 94305

(Received 26 January 2006; accepted 5 May 2006; published online 27 June 2006)

We propose an equi-energy (EE) sampling approach to study protein folding in the two-dimensional hydrophobic-hydrophilic (HP) lattice model. This approach enables efficient exploration of the global energy landscape and provides accurate estimates of the density of states, which then allows us to conduct a detailed study of the thermodynamics of HP protein folding, in particular, on the temperature dependence of the transition from folding to unfolding and on how sequence composition affects this phenomenon. With no extra cost, this approach also provides estimates on global energy minima and ground states. Without using any prior structural information of the protein the EE sampler is able to find the ground states that match the best known results in most benchmark cases. The numerical results demonstrate it as a powerful method to study lattice protein folding models. © *2006 American Institute of Physics*. [DOI: 10.1063/1.2208607]

I. INTRODUCTION

The prediction of protein structure from its primary sequence is a long-standing problem in biology.[1,2] The difficulty of this problem is due to the roughness of the energy landscape with a multitude of local energy minima separated by high barriers. Conventional Monte Carlo and molecular dynamics simulations tend to become trapped in local minima and are hence incapable of exploring the global energy surface. Even in simplified lattice models, the problem of finding the ground state of the protein is NP-complete.[3–5]

Many computation strategies have been proposed and extensively tested to address this difficulty, including Monte Carlo with minimization,[6] simulated annealing,[7] genetic algorithms,[8] multicanonical sampling,[9,10] evolutionary Monte Carlo,[11] human-guided search algorithms,[12] coredirected chain growth method,[13] pruned-enriched Rosenbluth method,[14] and sequential importance sampling with pilot-exploration resampling.[15]

Traditionally, the computational focus of the protein folding problem has been on finding the global minimal energy conformation. In this paper we take an alternative perspective where the aim is to sample the entire phase space. Compared with the traditional optimization approach, which neglects the conformations' entropic contributions, this sampling approach has the advantage of being able to estimate the thermodynamic quantities of interest (in addition to finding the ground state). We use our new sampling method, the equi-energy (EE) sampler,[16] to study the two-dimensional (2D) hydrophobic-hydrophilic (HP) model[17,18] in this paper. The key ingredient of the EE sampler is a new type of move called the equi-energy jump that aims to explore the phase space by moving directly between states with similar energy (see Fig. 1 for an illustration). It is motivated from the observation that for a given Boltzmann distribution $p(\mathbf{x}) \propto \exp(-h(\mathbf{x})/k_B T)$, if a sampler can move freely between any two states $\mathbf{x}$ and $\mathbf{y}$ with the same energy, i.e., $h(\mathbf{x})=h(\mathbf{y})$, then the problem of local trapping will be effectively eliminated.

Using the EE sampler, we estimate the density of states (of the phase space), which then allows us to investigate the thermodynamics of HP protein folding, in particular, (i) how the thermodynamic properties of protein folding change as the temperature varies, and (ii) what affects the temperature dependence. We find numerically that there appears to be in general a transition temperature associated with protein folding, where the HP protein experiences a sharp transition from unfolded states to orderly folded states (see Sec. IV). Furthermore, the EE sampler's ability to extensively explore the energy surface also leads to efficient search for the ground state. Without using any prior structural information of the protein the EE sampler is able to find the ground states that match the best known results in all but one benchmark case, where the next lowest energy level is reached.

It has been shown recently[19–21] that the pairwise additive hydrophobic contact energy in the HP model is not sufficient

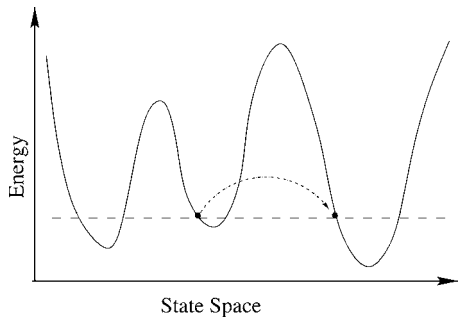

FIG. 1. Illustration of the equi-energy jump.

[a]Electronic mail: kou@stat.harvard.edu

to satisfy the cooperativity criterion, in particular, the calorimetric two-state constraint on folding/unfolding transitions. Despite this limitation of the HP model, the powerful sampling ability displayed by the EE sampler demonstrates its potential applicability to study the thermodynamics of general protein folding models. By simply adopting different energy functions, we expect the EE sampler to be a useful tool to evaluate the successes and limitations of protein folding models in capturing the behavior of real proteins.

The paper is organized as follows. Section II reviews the 2D HP model and introduces the EE sampler. Section III explains how to estimate the density of states using the EE sampler. Section IV studies the thermodynamic properties of HP protein folding. Section V focuses on finding the ground state. Section VI concludes the paper with discussion.

II. THE 2D HP MODEL AND THE EE SAMPLER

Among protein folding models, the HP model[17,18,22] is perhaps the most popular. The interest in this model arises from the realization that although simple, it does exhibit features of real protein folding.[23,24] A protein conformation in the 2D HP model is modeled as a 2D self-avoiding walk on a square lattice. The "amino acids" of the protein are simplified to only two types: a hydrophilic type (P-type) that favors interaction with water molecules, and a hydrophobic type (H-type) that does not interact well with water. Energies $\epsilon_{HH}=-1$, $\epsilon_{HP}=\epsilon_{PP}=0$ are assigned to interactions between noncovalently bound neighbors on the lattice. The total energy of a given conformation is simply the sum of energy contributions from the (noncovalently) interacting lattice neighbors. This energy assignment leads to the desirable feature that upon folding the hydrophobic residues typically form a compact core surrounded by a hydrophilic shell.[17,18]

The conformation of a length-k HP protein can thus be described by a vector $\mathbf{x}=(x_1,x_2,\ldots,x_k)$, where x_i is the lattice position of the ith residue of the protein. If we use $h(\mathbf{x})$ to denote the energy function, then the Boltzmann distribution is given by

$$\pi(\mathbf{x}) \propto \exp(-h(\mathbf{x})/T),$$

where T is the temperature. Sampling the Boltzmann distribution faces the major challenge of local energy traps. The EE sampler overcomes this difficulty by performing equi-energy jumps (Fig. 1) in the phase space. To do so, the EE sampler exploits two facts. First, at high temperature, where the Boltzmann distribution is relatively flat, it is easier to escape the energy traps. Second, the microcanonical ensembles, defined as the collection of conformations having the same energy $\{\mathbf{x}:h(\mathbf{x})=u\}$, are independent of temperature, which implies that if the microcanonical ensembles are constructed at a high temperature, they will remain valid at low temperatures.

In the EE sampler, a sequence of energy levels is introduced,

$$H_1 < H_2 < \cdots < H_K < 0,$$

such that H_1 is below the minimum energy $H_1 \leqslant \min_x h(x)$. Associated with the energy levels is a sequence of temperatures,

$$T_1 < T_2 < \cdots < T_K.$$

The EE sampler considers a population of K distributions, each indexed by a temperature and an energy truncation. The probability density function of the ith distribution π_i $(1 \leqslant i \leqslant K)$ is $\pi_i(x) \propto \exp(-h_i(x))$, where $h_i(x) = (1/T_i)\max(h(x),H_i)$. The energy truncation is used here to flatten the distribution for easier exploration. [The notation $\max(a,b)$ denotes the larger of a and b.]

The EE sampler begins from running a Markov chain $X^{(K)}$ targeting the highest order distribution π_K. The Markov chain state is updated by a mutation operator (described below shortly). After updating $X^{(K)}$ for a while (the burn-in period), the EE sampler starts constructing the Kth order microcanonical ensemble $D_u^{(K)}$ by grouping the samples according to their energy levels, i.e., $D_u^{(K)}$ consists of all the samples $X_n^{(K)}$ such that their energy $h(X_n^{(K)})=u$ $(u=H_1,H_1+1,\ldots,-1,0)$. This step is necessary because the equi-energy jump requires the knowledge of the microcanonical ensemble, which is not known *a priori*. After a fixed number of N iterations, the EE sampler starts the second highest order sampling chain $X^{(K-1)}$ targeting π_{K-1}, while keeps on running $X^{(K)}$ and updating $D_u^{(K)}$. The sampling chain $X^{(K-1)}$ is updated through two operations: the mutation (described below) and the equi-energy jump. At each update a coin is flipped; with probability $1-p_{EE}$, say, 90%, the current configuration $X_n^{(K-1)}$ undergoes a mutation to give the next state $X_{n+1}^{(K-1)}$, and with probability p_{EE}, say, 10%, $X_n^{(K-1)}$ goes through an equi-energy jump to yield $X_{n+1}^{(K-1)}$. In the equi-energy jump suppose $v=h(X_n^{(K-1)})$ is the energy of the current configuration; a state $\mathbf{y}$ chosen uniformly from the highest order microcanonical ensemble $D_v^{(K)}$ (in which all the conformations have energy v by construction) is then taken to be the next state $X_{n+1}^{(K-1)}$—the sampler thus jumps from $X_n^{(k-1)}$ to $\mathbf{y}$.

After updating chain $X^{(K-1)}$ for a while, the EE sampler starts the construction of the second highest order microcanonical ensemble $D_u^{(K-1)}$ in much the same way as the construction of $D_u^{(K)}$, i.e., grouping the samples according to their energy levels [$D_u^{(K-1)}$ consists of all the samples $X_n^{(K-1)}$ such that their energy $h(X_n^{(K-1)})=u$]. Once the chain $X^{(K-1)}$ has been running for N steps, the EE sampler starts $X^{(K-2)}$ targeting π_{K-2} while keeps on running $X^{(K-1)}$ and $X^{(K)}$ and updating $D_u^{(K-1)}$ and $D_u^{(K)}$. Like $X^{(K-1)}$, the sampling chain $X^{(K-2)}$ is updated by the mutation operator and the equi-energy jump with probabilities $1-p_{EE}$ and p_{EE}, respectively. In the equi-energy jump, a state $\mathbf{y}$ uniformly chosen from $D_{h(X_n^{(K-2)})}^{(K-1)}$, where $X_n^{(K-2)}$ is the current state, is taken to be the next state $X_{n+1}^{(K-2)}$. The EE sampler in this way successively steps down the energy and temperature ladder until the distribution π_1 is reached. Each chain $X^{(i)}$, $1 \leqslant i < K$, is updated by the equi-energy jump and the mutation. The microcanonical ensembles $D_u^{(i)}$ in each chain $X^{(i)}$ are constructed after a

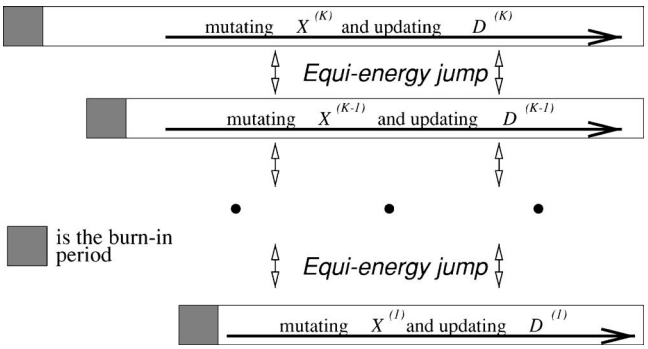

FIG. 2. Diagram of the EE sampler.

burn-in period, and are used by chain $X^{(i-1)}$ in the equi-energy jump. Figure 2 diagrams the sampling scheme.

The mutation operator (referred to above) governs how to go from one state to another to explore the phase space. In our implementation of the EE sampler we use the pull moves suggested by Ref. 12 as the mutation move set. The pull moves are local, reversible, and complete,[12] which makes them efficient for the mutation operation. They work as follows.

Consider the lattice points x_i and x_{i+1} occupied by the ith and $(i+1)$th residue (of the protein). Let L_1 and L_2 denote the other two lattice points that are adjacent to x_{i+1} [Fig. 3(a)]. If one of L_1 or L_2 is unoccupied, call it L, label the fourth lattice point in this minisquare as C [Fig. 3(b)]. If $C=x_{i-1}$ [i.e., occupied by the $(i-1)$th residue], then the pull move simply moves x_i to occupy L, generating a new configuration [Fig. 3(c)]. If both C and L are unoccupied, the pull move then operates by moving x_i to L, x_{i-1} to C, and moving x_{i-2} to where x_i used to be, x_{i-3} to where x_{i-1} used to be, …, x_{i-j} to where x_{i-j+2} used to be, until a new legal conformation is reached by the least number of lattice moves [Fig. 3(d)]. In the pull move described above the residues are pulled one by one in descending order. By symmetry the residue positions can also be pulled in ascending order in a pull move.

The pull moves also include end moves for the purpose of reversibility. Let L and C be two adjacent unoccupied lattice points such that L is adjacent to the first residue position x_1. The end move displaces x_1 to C, x_2 to L, and x_3 to the position used to be occupied by x_1, x_j to the position used to be occupied by x_{j-2}, etc., until a new legal configuration is reached by the least number of lattice moves [Fig. 3(e)]. The end move on the last residue position is similarly defined by symmetry.

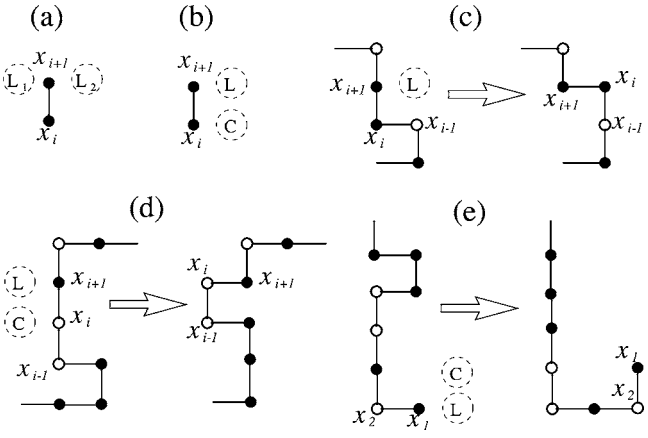

FIG. 3. The pull moves as the mutation move set.

During the sampling process, to mutate a given state, say, $X_n^{(i)}$, the EE sampler counts the number of pull moves [the three types shown in Figs. 3(c)–3(e)] of the current configuration. One of these possible moves is then chosen uniformly and applied to obtain another configuration $\mathbf{z}$. To maintain π_i as the invariant distribution after the mutation operation, $\mathbf{z}$ is accepted to be the next state $X_{n+1}^{(i)}$ with the Metropolis-Hastings[25,26] type probability

$$p_{\text{accept}} = \min\left(1, \frac{\pi_i(\mathbf{z})P(\mathbf{z} \to X_n^{(i)})}{\pi_i(X_n^{(i)})P(X_n^{(i)} \to \mathbf{z})}\right),$$

where $P(\mathbf{z} \to X_n^{(i)})$ = (number of pull moves from $\mathbf{z}$ to $X_n^{(i)}$)/(total number of pull moves of $\mathbf{z}$) is the probability of $\mathbf{z}$ mutating to $X_n^{(i)}$, and $P(X_n^{(i)} \to \mathbf{z})$, the probability of $X_n^{(i)}$ mutating to $\mathbf{z}$, is similarly defined. Since the calculation of the Metropolis-Hastings probability involves counting the total number of pull moves of both $X_n^{(i)}$ and $\mathbf{z}$, the computational complexity of a mutation move is roughly twice that of simply pull moving $X_n^{(i)}$ to $\mathbf{z}$.

The EE sampler has three user-choice parameter sets: the equi-energy jump probability p_{EE}, the energy levels H_1, H_2,…, and the temperature ladder $T_1, T_2, \ldots, T_K$. In our experience, the following choices appear to work well: take p_{EE} to be between 0.05 and 0.2, assign the energy levels through a rough geometric progression, and set the temperatures to be between 0.01 and 4 via a geometric progression. Ref. 16 provides more discussions about the parameter choice.

III. DENSITY OF STATES ESTIMATION

In the study of a statistical mechanical system the density of states $\Omega(u)$, defined as

$$\Omega(u) = \#\{\mathbf{x}: h(\mathbf{x}) = u\},$$

(whose logarithm is referred to as the microcanonical entropy) plays an important role, because many thermodynamic quantities can be directly calculated from the density of states.[27] (Throughout this paper, the notation $\#A$ denotes the total number of elements in set A.) As the construction of the microcanonical ensembles $D_u^{(i)}$ is an integral part of the EE sampler, it leads to a simple way to estimate the density of states. Under distribution π_i the probability $P_{\pi_i}(h(X)=u)$ of observing a state with energy u is given by

$$P_{\pi_i}(h(X) = u) = \frac{\Omega(u)e^{-\max(u,H_i)/T_i}}{\Sigma_v \Omega(v)e^{-\max(v,H_i)/T_i}}. \quad (1)$$

Since the density of states $\Omega(u)$ is common for all π_i, we can use the maximum likelihood method to combine all the microcanonical ensembles obtained from the different chains to estimate $\Omega(u)$. To do so, denote $m_u^i = \#D_u^{(i)}$, $m_\bullet^i = \Sigma_u m_u^i$, $m_u^\bullet = \Sigma_i m_u^i$, and $a_u^i = e^{-\max(u,H_i)/T_i}$ for notational convenience. Equation (1) leads to a multinomial distribution for the counts m_u^i,

$$(m^i_{H_1},\ldots,m^i_u,\ldots,m^i_0)\sim \text{multinomial}\left(m^i_\cdot;\frac{\Omega(H_1)a^i_{H_1}}{\Sigma_v\Omega(v)a^i_v},\ldots,\frac{\Omega(u)a^i_u}{\Sigma_v\Omega(v)a^i_v},\ldots,\frac{\Omega(0)a^i_0}{\Sigma_v\Omega(v)a^i_v}\right),\quad i=1,\ldots,K,$$

meaning that the joint probability distribution function of $(m^i_{H_1},\ldots,m^i_u,\ldots,m^i_0)$ is given by

$$\frac{(m^i_\cdot)!}{\Pi_u(m^i_u)!}\prod_u\left(\frac{\Omega(u)a^i_u}{\Sigma_v\Omega(v)a^i_v}\right)^{m^i_u}.$$

Correspondingly the likelihood function is

$$\text{lik}(\boldsymbol{\Omega})\propto\prod_i\prod_u\left(\frac{\Omega(u)a^i_u}{\Sigma_v\Omega(v)a^i_v}\right)^{m^i_u},$$

where the vector $\boldsymbol{\Omega}=(\Omega(H_1),\Omega(H_1+1),\ldots,\Omega(0))$. The maximum likelihood estimate $\hat{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega}$ is then defined as the maximizer of the likelihood function,

$$\begin{aligned}\hat{\boldsymbol{\Omega}}&=\arg\max_{\boldsymbol{\Omega}}\{\text{lik}(\boldsymbol{\Omega})\}\\&=\arg\max_{\boldsymbol{\Omega}}\{\log(\text{lik}(\boldsymbol{\Omega}))\}\\&=\arg\max_{\boldsymbol{\Omega}}\left\{\sum_u m^i_\cdot\log(\Omega(u))-\sum_i m^i_\cdot\log\left(\sum_v\Omega(v)a^i_v\right)\right\}.\end{aligned}$$

Since $\hat{\boldsymbol{\Omega}}$ maximizes the above expression, it must satisfy

$$\frac{\partial}{\partial\boldsymbol{\Omega}}\left(\sum_u m^i_\cdot\log(\Omega(u))-\sum_i m^i_\cdot\log\left(\sum_v\Omega(v)a^i_v\right)\right)\Bigg|_{\boldsymbol{\Omega}=\hat{\boldsymbol{\Omega}}}=0,$$

which leads to a set of equations for $\hat{\boldsymbol{\Omega}}$,

TABLE I. The normalized density of states estimated from the EE sampler (plus and minus twice the standard error) compared with the actual value for the length-20 protein HPHPPHHPHPPHPHHPPHPH.

Energy	Estimated density of states	Actual value
−9	$(4.738\pm1.403)\times10^{-8}$	4.774×10^{-8}
−8	$(1.162\pm0.113)\times10^{-6}$	1.146×10^{-6}
−7	$(1.452\pm0.075)\times10^{-5}$	1.425×10^{-5}
−6	$(1.248\pm0.046)\times10^{-4}$	1.237×10^{-4}
−5	$(9.309\pm0.305)\times10^{-4}$	9.200×10^{-4}
−4	$(6.245\pm0.146)\times10^{-3}$	6.183×10^{-3}
−3	$(3.554\pm0.053)\times10^{-2}$	3.514×10^{-2}
−2	$(1.499\pm0.011)\times10^{-1}$	1.489×10^{-1}
−1	$(3.779\pm0.016)\times10^{-1}$	3.779×10^{-1}
0	$(4.294\pm0.018)\times10^{-1}$	4.309×10^{-1}

$$\frac{m^i_u}{\hat{\Omega}(u)}-\sum_i\frac{m^i_\cdot a^i_u}{\Sigma_v\hat{\Omega}(v)a^i_v}=0\quad\text{for all } u.\tag{2}$$

Equation (2) can be used to compute $\hat{\Omega}(u)$ through a simple iteration

$$\hat{\Omega}(u)=m^i_u\Big/\left(\sum_i\frac{m^i_\cdot a^i_u}{\Sigma_v\hat{\Omega}(v)a^i_v}\right).\tag{3}$$

To use this expression, we note that since $\Omega(u)$ is specified up to a scale change [see Eq. (1)], to estimate the relative values we can set without loss of generality $\Omega(u_0)=1$ for some u_0.

To test our strategy to estimate the density of states we apply the EE sampler on a length-20 protein HPHPPHHPH-PPHPHHPPHPH, which is sequence 1 in Ref. 8. For this

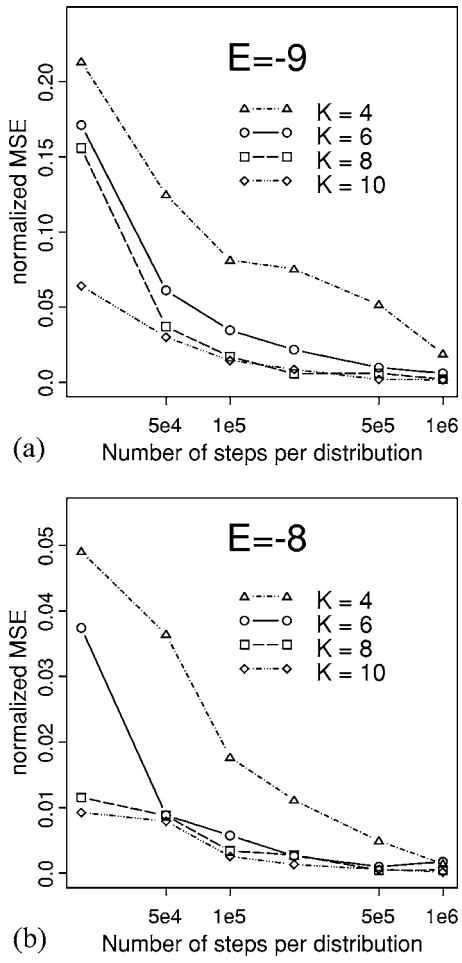

FIG. 4. The normalized mean square error of $\hat{\Omega}(u)$ for various combinations of K, the number of distributions employed, and N, the number of Monte Carlo steps per distribution. (a) The accuracy for estimating the density of states $\Omega(u)$ at the lowest energy of −9. (b) The accuracy for estimating $\Omega(u)$ at the second lowest energy of −8.

TABLE II. The normalized density of states estimated from the EE sampler (plus and minus twice the standard error) for the length-64 protein HHHHHHHHHHHHPHPHPPHHPPHHPPHPPHHPPHHPPHPPHHPPHPHPHHHHHHHHHHHH.

Energy	Estimated density of states	Energy	Estimated density of states
−42	$(1.021\pm0.726)\times10^{-25}$	−20	$(3.205\pm0.145)\times10^{-7}$
−41	$(3.816\pm7.035)\times10^{-25}$	−19	$(1.224\pm0.050)\times10^{-6}$
−40	$(1.008\pm0.474)\times10^{-23}$	−18	$(4.478\pm0.156)\times10^{-6}$
−39	$(2.041\pm0.980)\times10^{-22}$	−17	$(1.564\pm0.047)\times10^{-5}$
−38	$(3.105\pm1.524)\times10^{-21}$	−16	$(5.215\pm0.132)\times10^{-5}$
−37	$(2.988\pm1.215)\times10^{-20}$	−15	$(1.639\pm0.034)\times10^{-4}$
−36	$(2.832\pm1.090)\times10^{-19}$	−14	$(4.826\pm0.084)\times10^{-4}$
−35	$(2.488\pm0.902)\times10^{-18}$	−13	$(1.321\pm0.018)\times10^{-3}$
−34	$(2.025\pm0.684)\times10^{-17}$	−12	$(3.348\pm0.037)\times10^{-3}$
−33	$(1.529\pm0.482)\times10^{-16}$	−11	$(7.791\pm0.066)\times10^{-3}$
−32	$(1.078\pm0.310)\times10^{-15}$	−10	$(1.664\pm0.011)\times10^{-2}$
−31	$(7.178\pm1.823)\times10^{-15}$	−9	$(3.242\pm0.016)\times10^{-2}$
−30	$(4.514\pm1.004)\times10^{-14}$	−8	$(5.724\pm0.023)\times10^{-2}$
−29	$(2.683\pm0.504)\times10^{-13}$	−7	$(9.107\pm0.028)\times10^{-2}$
−28	$(1.519\pm0.239)\times10^{-12}$	−6	$(1.293\pm0.003)\times10^{-1}$
−27	$(8.216\pm1.085)\times10^{-12}$	−5	$(1.615\pm0.002)\times10^{-1}$
−26	$(4.239\pm0.470)\times10^{-11}$	−4	$(1.735\pm0.002)\times10^{-1}$
−25	$(2.091\pm0.198)\times10^{-10}$	−3	$(1.543\pm0.003)\times10^{-1}$
−24	$(9.879\pm0.795)\times10^{-10}$	−2	$(1.068\pm0.004)\times10^{-1}$
−23	$(4.474\pm0.309)\times10^{-9}$	−1	$(5.106\pm0.027)\times10^{-2}$
−22	$(1.937\pm0.116)\times10^{-8}$	0	$(1.305\pm0.015)\times10^{-2}$
−21	$(8.048\pm0.423)\times10^{-8}$		

protein there are 83 779 155 possible conformations with energies ranging from 0 to −9 (known through exhaustive enumeration). Table I presents the estimated and exact (normalized) density of states at various energy levels. The estimates are based on the average of ten independent runs, each using six ($K=6$) distributions and $N=200\,000$ steps per distribution, where the energy truncation levels were set by a geometric progression between −9.5 and −1.5, the temperatures were assigned by a geometric progression between 0.15 and 2, and the equi-energy jump probability p_{EE} was set to 10%.

It is clear that the method yielded very accurate estimates of the density of states at each energy level even though we have sampled only a small fraction of the population of conformations. The equi-energy jump is important for the success of the method: if we eliminate these jumps, the results were very poor in estimating the density of states at the low energy end; for example, the estimated density of state at $E=-9$ becomes $(3.623\pm1.329)\times10^{-10}$, which is two orders of magnitude away from the exact value.

The EE sampler employs K distributions. Intuitively, the estimation accuracy increases with K and N, the number of Monte Carlo steps per distribution. Figure 4 shows this dependence; it plots, for the two lowest energy levels −9 and −8, the normalized mean square error of the estimate $\hat{\Omega}(u)$, as K and N vary; here the normalized mean square error, measuring the estimation accuracy, is defined as the expectation of $([\hat{\Omega}(u)-\Omega(u)]/\Omega(u))^2$, where $\hat{\Omega}(u)$ was obtained from the average of ten independent runs of the EE sampler.[28] Since estimating the density of states at the lowest energy levels is most challenging (as seen in Table I), Fig. 4 demonstrates the efficiency of the EE sampler in that for K as small as 6, and N as small as 200 000, the EE sampler already produces very accurate estimates.

We next apply the EE sampler to estimate the density of states of a length-64 protein, which is sequence 8 in Ref. 8 (see Table II for its composition). For this protein, exhaustive enumeration is no longer feasible even with a supercomputer. Table II reports the EE sampler estimated density of states based on the average of 15 independent runs. Each run employs $K=35$ distributions and $N=2\times10^6$ steps per distribution. The energy truncation levels were set by a geometric progression between −42.5 and −3, the temperatures were assigned by a geometric progression between 0.02 and 3, and the equi-energy jump probability p_{EE} was set to 10%.

Table II reveals two interesting features of this energy landscape. First, unlike the short protein in Table I, the energy landscape of this longer chain is no longer dominated by states with zero energy. The mode of the density of states is now at $E=-4$. This interesting fact is a consequence of the length of the protein. With so many hydrophobic residues, a few of them end up being topological neighbors in most conformations. The second feature, which is more important, is that at lower energies, the difference in the density of states becomes larger. For example, the density of states for $E=-42$ is about two orders of magnitude smaller than that for $E=-40$. This finding suggests the potential limitation of the traditional approach of focusing on finding the ground state under the HP model. If there is such a large difference in density of states as indicated by Table II, then for the length-64 protein to prefer to be in a state of $E=-42$ over $E=-40$, the equilibrium temperature must be low, T

<0.436 indeed. This means that only searching for the minimal energy conformation is not necessarily adequate when characterizing the protein's behavior at modest temperature, say, $T=0.5$, under the HP model. We will discuss these issues further in the subsequent sections.

IV. THERMODYNAMIC PROPERTIES OF PROTEIN FOLDING

In this section we study the thermodynamics of HP protein folding, in particular, its temperature dependence. The density of states estimates from the EE sampler play a pivotal role here, because they provide a straightforward means to calculate any Boltzmann average of interest. Suppose for a state function g we want to estimate the Boltzmann average $\langle g\rangle_T$ at temperature T. We can write

$$\langle g\rangle_T = \frac{\Sigma_u \Omega(u) e^{-u/T} v_g(u)}{\Sigma_u \Omega(u) e^{-u/T}}, \tag{4}$$

where $v_g(u)$ is the microcanonical average of g on the microcanonical ensemble $\{\mathbf{x}:h(\mathbf{x})=u\}$. Therefore, if we estimate $v_g(u)$ by the combined sample average over $D_u^{(i)}$ $(i=1,2,\ldots,K)$,

$$v_g(u) \leftarrow \frac{\Sigma_i \Sigma_{\mathbf{x}\in D_u^{(i)}} g(\mathbf{x})}{\Sigma_i \Sigma_{\mathbf{x}\in D_u^{(i)}} \# D_u^{(i)}},$$

and substitute $\Omega(u)$ by its estimate $\hat{\Omega}(u)$, Eq. (4) then leads to a simple estimate of the Boltzmann average at any temperature T.

With this estimation method we study under the HP model how the folding behavior changes from high temperature, where the conformational distribution is dominated by the entropy term and the protein is likely to be in an unfolded state with high energy, to low temperature, where the conformation is likely to be compactly folded structures with low energy. Although the HP model has been shown to be insufficient to capture the cooperativity of real protein folding transitions,[19–21] we find it still instructive to demonstrate the EE sampler's ability to explore temperature dependent thermodynamic quantities. The application of the equi-energy sampler to a more physically realistic model can be accomplished by adopting a more physically realistic energy function.

We first use two statistics, the minimum box size (BOXSIZE) and the end-to-end distance, to measure the extent the HP protein has folded. BOXSIZE is defined as the area of the smallest possible rectangular region on the lattice containing all the amino acid positions in the conformation, whereas the end-to-end distance is the Euclidean distance between the two ends of the conformation. Intuitively, the averages of both statistics should increase with temperature. For the length-20 protein of Table I, Figs. 5(a) and 5(b) plot the estimated Boltzmann averages of BOXSIZE and the end-to-end distance as a function of temperature, which are available from the ten independent runs described in the previous section. The EE sampler is seen to very accurately estimate the Boltzmann values, and as expected both statistics increase with temperature.

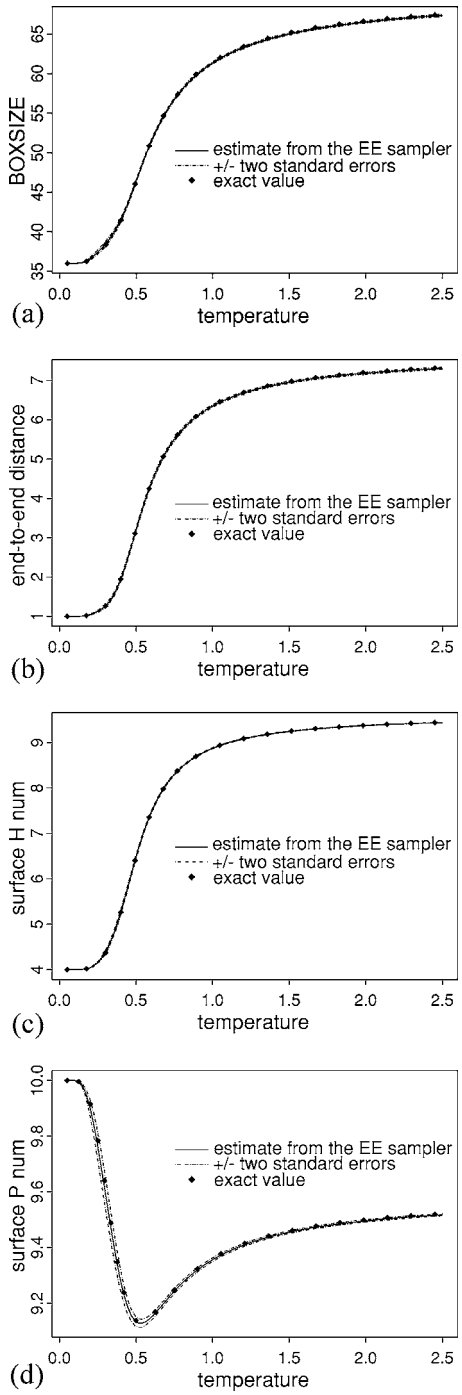

FIG. 5. The Boltzmann averages of (a) BOXSIZE, (b) end-to-end distance, (c) surface-H-number, and (d) surface-P-number at different temperatures of the length-20 protein.

Figures 5(a) and 5(b) also reveal that there is a rather sharp transition from order (folded state) to disorder (unfolded state) between $T=0.25$ and $T=1$, with an inversion point around $T=0.5$. The apparent transition raises an interesting question. Does the environment in which proteins live more closely resemble the $T=0.25$ scenario, or that of $T=0.5$ or even $T=1$? We therefore seek to correspond the room temperature to the unitless temperature considered here in the HP model. To find an overall measure of the strength of the hydrophobic interaction, we used the Miyazawa and Jernigan energies in Ref. 29, where a list of all the pairwise

interaction energies of the various amino acid residues (constructed statistically from databases of protein native structure[30]) is provided. We divided the residues into hydrophobic and hydrophilic ones. Averaging all of the hydrophobic-hydrophobic (HH) interaction energies, we found a value of -5.01 ± 0.34 (in RT units). Averaging all of the hydrophilic-hydrophilic (PP) and hydrophilic-hydrophobic (HP) interaction energies gave a value of -2.52 ± 0.14 (in RT units). Taking the difference, we thus found that the energy gap between the HH and HP/PP contact is -2.49 ± 0.36 (in RT units), which is -1494 ± 216 cal/mol under the conversion[29] of $RT=600$ cal/mol. Since in the HP model the energy gap between the HH contact and the HP/PP contact is -1 ($\epsilon_{\mathrm{HP}}=\epsilon_{\mathrm{PP}}=0$, $\epsilon_{\mathrm{HH}}=-1$), matching up

$$\frac{-1}{T_{\mathrm{room}}}=\frac{-1494\pm216\ \mathrm{cal/mol}}{k_B 293\ \mathrm{K}},$$

and using the Boltzmann factor, we found that room temperature of 293 K corresponds to the unitless temperature of $T_{\mathrm{room}}=0.390\pm0.056$ in the HP model. It should be noted that, rigorously speaking, the hydrophobic interaction is not temperature independent.[31] Nevertheless, the above correspondence between the unitless temperature in the HP model and the real room temperature serves as a rough guideline.

With this rough correspondence of $T_{\mathrm{room}}=0.390\pm0.056$, Fig. 5 suggests that a HP protein like the length-20 one does not necessarily always assume the minimum energy conformation at room temperature; it might have a nontrivial probability of being in high-energy, unfolded states. If this is the case, then conventional wisdom of focusing on finding the minimum energy conformation might only reflect part of the picture of the HP protein folding model. Considering the thermodynamics (such as the equilibrium statistics) offers a more comprehensive understanding.

We use two more statistics to further study this apparent transition behavior: the surface hydrophobic residue number (surface-H-number) and the surface hydrophilic residue number (surface-P-number). The surface-H-number of a conformation is defined as the number of hydrophobic residues that have direct contact with the outside (as opposed to being imbedded inside). The surface-P-number is defined similarly. Figures 5(c) and 5(d) show the estimated Boltzmann average of both statistics at various temperatures. Both plots confirm the sharp transition around $T=0.5$. The monotonic increase of average surface-H-number with temperature conforms with the picture that as temperature gets lower the hydrophobic residues are forced to stay inside to minimize the energy. The V-shaped curve of average surface-P-number, on the other hand, suggests a more interesting picture. At very high temperature, the HP proteins are essentially unfolded with a large number of hydrophilic residues having contact with outside; as the temperature drops the protein starts to be partly folded, and consequently some hydrophilic residues happen to be folded inside, resulting in a drop of average surface-P-number; as the temperature drops even lower the protein becomes fully folded, and in order to minimize the energy the protein has to squeeze out the hydrophilic residues to the surface to make room inside for the hydrophobic residues, which results in the increase of average surface-P-number once the temperature is below certain threshold. As the dip of its V curve corresponds nicely to the transition point at around $T=0.5$, the surface-P-number appears to be a good indicating statistic for the transition behavior under the HP model.

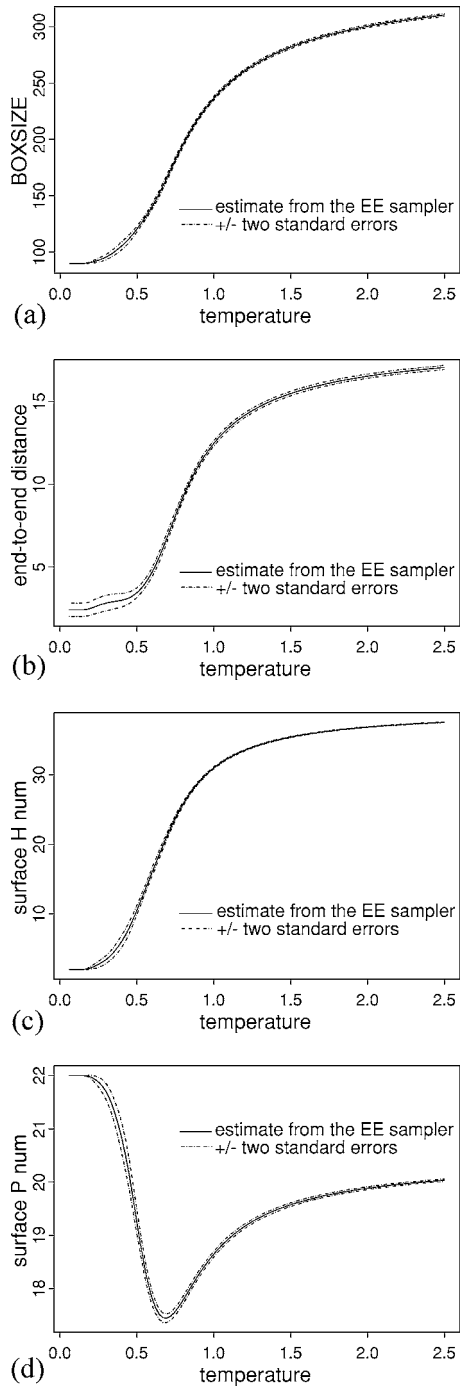

FIG. 6. The Boltzmann averages of (a) BOXSIZE, (b) end-to-end distance, (c) surface-H-number, and (d) surface-P-number at different temperatures of the length-64 protein.

Figure 6 shows the estimated Boltzmann averages of the four statistics for the length-64 protein of Table II, obtained from the 15 independent runs of Sec. III. Similar transition behavior is observed with the transition temperature around $T=0.68$.

TABLE III. The nine length-50 sequences together with their hydrophobic residue percentages, minimum energies, and the apparent transition temperatures.

Sequence code	Sequence	H%	Minimum energy	Apparent transition temperature
Seq50a	HPPPPPPPHHPPPPPPPPPPPPPPPPPHHPPPPPPHPHHPPHPPPPPHP	20%	−6	<0.15
Seq50b	PHPPPHPPPPPPPPHPPPPPPPPHPPPPHPPHPHHPPPPPPHPPPPPPHPH	20%	−8	<0.20
Seq50c	PHPPHPPPPHPPHPHHHPPHHPHPHHPPPHHPHPHPPPHHHPPPPPPPPH	40%	−16	≈0.22
Seq50d	PPHPHPPPHPHHPPHPPHPPHPPHPPHPHHPHHPHHPPPPHHPHPPPPH	40%	−18	≈0.25
Seq50e	PPPPHHHHHHHHPHPHPHHPHHHPHPPPHHPHHHHHPHPHPPPPHHH PHHP	60%	−23	0.43
Seq50f	HPHPHPHPHPHHHHHHPHPHHHPHPHPHHHPHHPPHPHPPPHHHPP PHHP	60%	−23	0.35
Seq50g	HPHHHPHHHHHHPPHHPHPPHHHHHPHHHPPPPHHHHPPPPHPPP HHPHP	60%	−24	0.43
Seq50h	PHHHHPHHHHHHHHHHHHPHHHPHHPHHHHHHHPHHPHHH HHHHPHHPHP	80%	−33	0.77
Seq50i	HPHHPHHHHHHPHHPHHHHHHHHHHHHHHHPHPPHHHHHHH PPPHHHHHH	80%	−34	0.90

The study of the length-20 and 64 proteins raises a question. What affects the transition temperature if there is one? One candidate could be the proportion of hydrophobic residues in the protein sequence versus that of hydrophilic ones. We thus generated nine length-50 proteins with different proportions of hydrophobic residues, where the hydrophobic residues were placed randomly along the sequence. Table III shows the nine sequences, and Fig. 7 for each of them plots the surface-P-number (the indicating statistic) versus the temperature. Each panel of Fig. 7 was obtained by 15 inde-

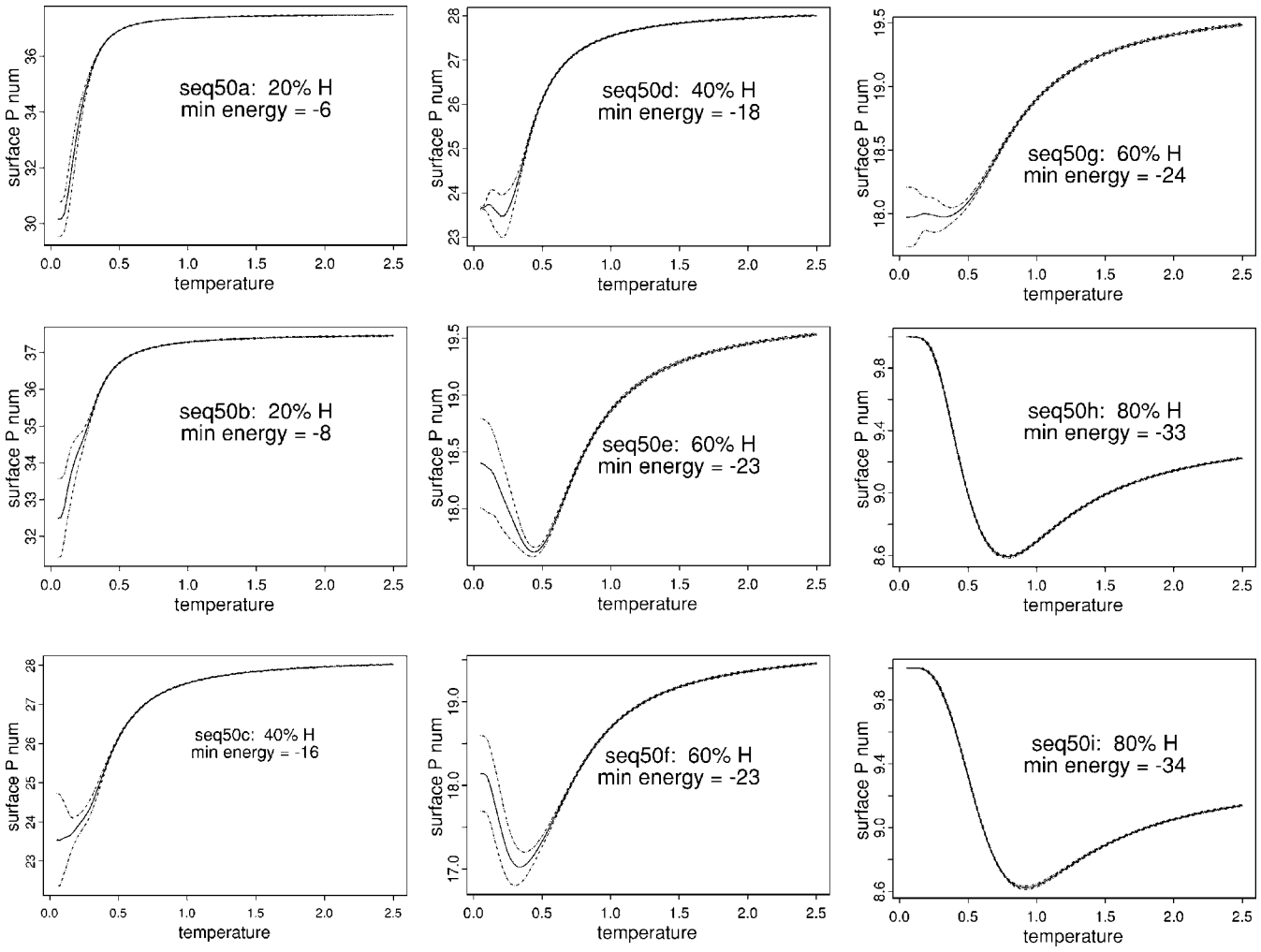

FIG. 7. The Boltzmann average (plus and minus two standard errors) of the surface-P-number of the nine length-50 sequences at different temperatures.

pendent runs of the EE sampler, each run employing 2×10^6 steps per distribution. Table III also lists the minimum energy and the apparent transition temperature of each sequence.

Three general features are observed from Table III and Fig. 7 for the HP model.

(a) While the V-shaped curve of the surface-P-number is more clearly seen for sequences with high proportions of hydrophobic residues, the transition behavior appears to hold in general.
(b) For a HP protein of fixed length, increasing the number of hydrophobic residues raises the transition temperature.
(c) It appears that in general for a HP protein with fixed length and fixed number of hydrophobic residues the lower the minimum energy the higher the transition temperature.

The possible explanation for observation (b) is that as the number of hydrophobic residues increase, the number of HH bonds (formed by HH neighbor pairs) in general also increases, which implies that the transition temperature from order to disorder has to be higher to break the increased number of HH bonds. Observation (c) could be explained similarly. The lower the minimum energy, the greater the number of HH bonds that can be formed, and consequently the higher the temperature is needed to make the transition from folded to unfolded.

Under the HP model, observations (a)–(c) could possibly lead to a connection between the transition behavior observed here, where the length-20, 64, and 50 proteins might not always assume their minimum energy at room temperature (as the room temperature appears not much lower than the transition temperature), and the hypothesis that in live cells a protein tends to fold quickly to its native state. Proteins in live cells tend to be quite long having many hydrophobic residues that could form many HH bonds, which in turn makes the transition temperature high. If the transition temperature is high enough (e.g., $T=1.5$ or 2 in the current temperature scale), then essentially at room temperature $T_{\text{room}}=0.390\pm0.056$, a protein may nearly always assume the minimum energy conformation. On top of this it is possible that in the evolution process those proteins that have transition temperature close to room temperature had gone extinct due to their instability, and only those having high enough transition temperature survived. More research beyond the HP model is needed to investigate this plausibility. But it is worth emphasizing that it is the EE sampler's extensive sampling ability that enables us to explore this phenomenon numerically.

V. GROUND STATES

The EE sampler is seen to be a powerful sampling algorithm. Its strength of globally exploring the energy landscape also makes it a capable optimization tool. In this section, we apply the EE sampler to nine benchmark HP sequences that are widely used in the literature to test search algorithms for ground states. Table IV lists the nine sequences. Table V summarizes the EE sampler's performance together with that of other methods reported in the literature, including genetic algorithms (GA),[8] evolutionary Monte Carlo (EMC),[11] pruned-enriched Rosenbluth method (PERM),[14] sequential importance sampling with pilot-exploration resampling (SISPER),[15] and human-guided search (HuGS).[12] The EE sampler achieves the best known result for all but one sequence. The parameter settings of the EE sampler for the four longest sequences are reported in Table VI.

Four sequences are particularly worth commenting. (i) Seq64 (the length-64 one) has been marked in Refs. 11, 14, and 15 as a very difficult one, and without imposing secondary structure GA, EMC, PERM, and SISPER were not able

TABLE IV. The nine benchmark sequences. Seq20 to seq85 are taken from Ref. 11. Seq100a and seq100b are taken from Ref. 32.

Sequence code	Length	Sequence
Seq20	20	HPHPPHHPHPPHPHHPPHPH
Seq36	36	PPPHHPPHHPPPPPHHHHHHHPPHHPPPPHHPPHP
Seq48	48	PPHPPHHPPHHPPPPPHHHHHHHHHHPPPPPPHHPPHHPPH PPHHHHH
Seq50	50	HHPHPHPHPHHHHPHPPPHPPPHPPPPHPPPHPPPHPHHHHP HPHPHPHH
Seq60	60	PPHHHPHHHHHHHHPPPHHHHHHHHHHPHPPPHHHHHHH HHHHPPPPHHHHHHPHHPHP
Seq64	64	HHHHHHHHHHHHPHPHPPHHPPHHPPHPPHHPPHHPPHP PHHPPHHPPHPHPHHHHHHHHHHH
Seq85	85	HHHHPPPPHHHHHHHHHHHHPPPPPPHHHHHHHHHHHHP PPHHHHHHHHHHHHPPPHHHHHHHHHHPPPHPPHHPP HHPPHPH
Seq100a	100	PPPPPPHPHHPPPPHHHPHHHHHPHHPPPPHHPPHHPHH HHHPHHHHHHHHHHPHHPHHHHHHHPPPPPPPPHHHHH HHPPHPHHHPPPPPHPHH
Sq100b	100	PPPHHPPHHHHPPHHHPHPHHPHHHHPPPPPPPHHHH HHPPHHHHHHPPPPPPPPHPHHPHHHHHHHHHHPPHH HPHHPHPPHPHHHPPPPPHHH

TABLE V. The performance of the EE sampler in finding the ground states compared with that of GA, EMC, PERM, SISPER, and HuGS. Columns 3–6 are adopted from Ref. 15. Column 7 is adopted from Ref. 12. Column 2 reports the lowest energy achieved by the EE sampler in 15 independent runs. The parameter settings of the EE sampler for the four longest sequences are given in Table VI.

Sequence code	EE	GA	EMC	PERM	SISPER	HuGS
Seq20	−9	−9	−9	−9	−9	
Seq36	−14	−14	−14	−14	−14	
Seq48	−23	−22	−23	−23	−23	
Seq50	−21	−21	−21	−21	−21	
Seq60	−36	−34	−35	−36	−36	
Seq64	−42	−37	−39	−40	−39	−42
Seq85	−53				−52	−53
Seq100a	−48			−47	−48	−48
Seq100b	−49			−48	−49	−50

TABLE VI. The parameter settings of the EE sampler for the four longest sequences. Both the temperatures and the energy truncation levels are set by a geometric progressing within the range shown.

Sequence code	Energy truncation range	Temperature range	Number of distributions involved	Steps per distribution	EE jump probability
Seq64	[−42.5, −3]	[0.02, 3]	35	2 000 000	10%
Seq85	[−53.5,−4.5]	[0.02, 3]	45	2 000 000	5%
Seq100a	[−48.5,−3.5]	[0.02,3.5]	35	3 500 000	5%
Seq100b	[−50.5, −3]	[0.02,3.7]	35	3 500 000	5%

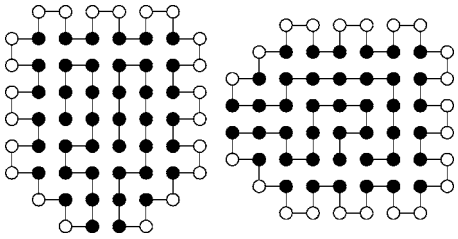

FIG. 8. Two conformations with energy of −42 for seq64 (the length-64 sequence) found by the EE sampler.

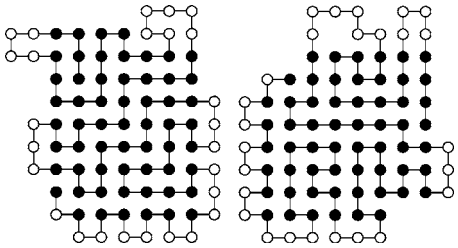

FIG. 9. Two conformations with energy of −53 for seq85 (the length-85 sequence) found by the EE sampler.

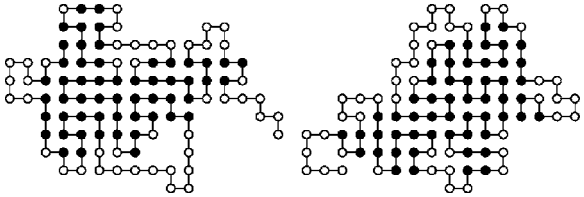

FIG. 10. Two conformations with energy of −48 for seq100a (the first length-100 sequence) found by the EE sampler.

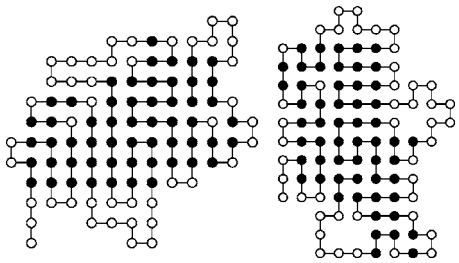

FIG. 11. Two conformations with energy −49 for seq100b (the second length-100 sequence) found by the EE sampler.

to reach the ground energy of −42. The EE sampler without using any structural information is able to not only fold the sequence to conformations with energy of −42, two of which are shown in Fig. 8, but also estimate the density of states as shown in Table II. (ii) Seq85 (the length-85 one) was introduced in Ref. 33, where the authors appeared to construct the sequence with ground energy of −52 in mind. But HuGS and the EE sampler are able to find conformations with energy of −53. Two such conformations found by the EE sampler are shown in Fig. 9. (iii) The minimum energy of seq100a (the first length-100 sequence) found by PERM is −47. The EE sampler, SISPER, and HuGS found conformations with energy of −48. Two such conformations found by the EE sampler are shown in Fig. 10. (iv) Seq100b (the second length-100 sequence) is the only sequence for which the minimum energy of −49 obtained by the EE sampler does not reach the best known result of −50. Two energy −49 conformations found by the EE sampler are shown in Fig. 11.

VI. DISCUSSION

With its extensive capability to explore the energy landscape, the EE sampler is seen as a powerful tool for not only finding the ground states but also providing efficient estimates of the density of states that allow subsequent computation of the statistical mechanical properties of protein folding. In particular, in addition to achieving the best known results for ground-state search in most cases, the numerical results from the EE sampler reveal the apparent transition phenomenon from disordered unfolding to orderly folding (associated with a transition temperature). This broader perspective of HP protein folding is manifested by the nine sequences of length 50 studied in Table III, where only two out of the nine have transition temperature significantly higher than room temperature (roughly $T_{\text{room}}=0.390\pm0.056$). For the majority seven sequences one cannot ignore the entropy term and must rely on sampling instead of optimization to study their folding behavior. Such detailed information is available in our study only because of the ability of the EE sampler to estimate density of states efficiently.

In this paper we have focused on the 2D HP model. But we expect that with appropriate generalization and enhancement, the method can be applied to study general three-dimensional (3D) lattice and off-lattice protein folding models, since previously the EE sampler has been successfully applied to problems of statistical inference, statistical mechanical calculation, and sequence alignment in computational biology.[16] As the equi-energy jump is a key step in the

EE sampler, we conclude this paper by a remark on its implementation. In Sec. II we showed that for the ith chain the equi-energy jump step could be performed by jumping from the current state of $X_n^{(i)}$ to a state uniformly chosen from the microcanonical ensemble $D^{(i+1)}_{h(X_n^{(i)})}$ constructed from the $(i+1)$th chain $X^{(i+1)}$. A possible generalization of this implementation is to allow $X_n^{(i)}$ to jump to a state uniformly chosen from the combined microcanonical ensemble of $\cup_{j=i+1}^{K} D^{(j)}_{h(X_n^{(i)})}$, which uses information from not only $X^{(i+1)}$ but also the other higher order chains $X^{(i+2)}, \ldots, X^{(K)}$ as well.

ACKNOWLEDGMENTS

One of the authors (S.C.K.) acknowledges support from NSF Grant No. DMS-02-04674, NSF Career Award, and NIH Grant No. R01HG02518. Another author (W.H.W.) is supported in part by NSF Grant No. DMS-0505732 and NIH grant P20-CA096470.

[1] M. Sela, F. H. White, and C. B. Anfinsen, Science **125**, 691 (1957).
[2] *Protein Folding*, edited by T. Creighton (Freeman, New York, 1993).
[3] R. Unger and J. Moult, Bull. Math. Biol. **55**, 1183 (1993).
[4] B. Berger and T. Leighton, J. Comput. Biol. **5**, 27 (1998).
[5] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, and M. Yannakakis, J. Comput. Biol. **5**, 423 (1998).
[6] Z. Li and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **84**, 6611 (1987).
[7] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, Science **220**, 671 (1983).
[8] R. Unger and J. Moult, J. Mol. Biol. **231**, 75 (1993).
[9] U. Hansmann and Y. Okamoto, J. Comput. Chem. **14**, 1333 (1993).
[10] U. Hansmann and Y. Okamoto, Curr. Opin. Struct. Biol. **9**, 177 (1999).
[11] F. Liang and W. H. Wong, J. Chem. Phys. **115**, 3374 (2001).
[12] N. Lesh, M. Mitzenmacher, and S. Whitesides, Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology, 2003, p. 188.
[13] T. C. Beutler and K. A. Dill, Protein Sci. **5**, 2037 (1996).
[14] U. Bastolla, H. Frauenkron, E. Gerstner, P. Grassberger, and W. Nadler, Proteins: Struct., Funct., Genet. **32**, 52 (1998).
[15] J. L. Zhang and J. S. Liu, J. Chem. Phys. **117**, 3492 (2002).
[16] S. C. Kou, Q. Zhou, and W. H. Wong, Ann. Stat. (to be published).
[17] K. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[18] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan, Protein Sci. **4**, 561 (1995).
[19] H. S. Chan, Proteins: Struct., Funct., Genet. **40**, 543 (2000).
[20] H. Kaya and H. S. Chan, Phys. Rev. Lett. **85**, 4823 (2000).
[21] H. S. Chan, S. Shimizu, and H. Kaya, Methods Enzymol. **380**, 350 (2004).
[22] K. A. Dill, Biochemistry **24**, 1501 (1985).
[23] K. Lau and K. Dill, Proc. Natl. Acad. Sci. U.S.A. **87**, 638 (1990).
[24] G. M. Crippen, Biochemistry **30**, 4232 (1991).
[25] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[26] W. K. Hastings, Biometrika **57**, 97 (1970).
[27] F. Mandl, *Statistical Physics* (Wiley, New York, 1988).
[28] The normalized mean square error is decomposed into a square bias term and a variance term. We estimated each term based on the ten independent runs.
[29] S. Miyazawa and R. Jernigan, J. Mol. Biol. **256**, 623 (1996).
[30] H. S. Chan, *Encyclopedia of Life Science* (Nature, London, UK, 2001).
[31] M. S. Moghaddam, S. Shimizu, and H. S. Chan, J. Am. Chem. Soc. **127**, 303 (2005).
[32] R. Ramakrishnan, B. Ramachandran, and J. F. Pekny, J. Chem. Phys. **106**, 2418 (1997).
[33] R. König and T. Dandekar, BioSystems **50**, 17 (1999).