Finding multiple minimum-energy conformations of the hydrophobic-polar protein model via multidomain sampling

Wei Tang and Qing Zhou

Phys. Rev. E **86**, 031909 — Published 11 September 2012

DOI: 10.1103/PhysRevE.86.031909

Finding Multiple Minimum-Energy Conformations of the Hydrophobic-Polar Protein Model via Multidomain Sampling

Wei Tang[1] and Qing Zhou[2*]
[1] Department of Materials Science and Engineering and [2] Department of Statistics,
University of California, Los Angeles, CA 90095, USA

August 1, 2012

Abstract

We demonstrate the efficiency of the multidomain sampler (MDS) in finding multiple distinct global minima and low-energy local minima in the hydrophobic-polar (HP) lattice protein model. Extending the idea of partitioning energy space in the Wang-Landau algorithm, our approach introduces an additional partitioning scheme to divide the protein conformation space into local basins of attraction. This double-partitioning design is very powerful in guiding the sampler to visit the basins of unexplored local minima. An H-residue subchain distance is used to merge the basins of similar local minima into one domain, which increases the diversity among identified minimum-energy conformations. Moreover, a visit-enhancement factor is introduced for long protein chains to facilitate jumps between basins. Results on three benchmark protein sequences reveal that our approach is capable of finding multiple global minima and hundreds of low-energy local minima of great diversity.

1 Introduction

The function of a protein not only depends on its amino acid sequence, but also strongly relies on its 3D spatial conformation. However, prediction of protein spatial structures from a given amino acid sequence has been a challenging computational problem, which is still under intensive investigation. Direct physical modeling, like molecular dynamics (MD), is powerful in studying folding kinetics and transition states, but is largely limited to short protein chains and fast folders (short time scale), due to the vast amount of computation required to construct the atomistic trajectories [1]. On the other hand, the rough energy landscape of protein folding structures also poses additional difficulty in search of the native state (energy global minimum) of a protein. Monte Carlo simulation based on all-atom models is able to partially relax the issue of short time scale in MD simulation [2, 3], because it does not need to follow the atomistic trajectories

*To whom correspondence should be addressed (email: zhou@stat.ucla.edu).

exactly. To further simplify the problem, coarse-grain lattice protein models are frequently used. The hydrophobic-polar (HP) model [4] is one of the most popular lattice protein models [5–9], as it greatly simplifies the amino acid representation and interaction. In this model, amino acids are abstracted as hydrophobic (H) or polar (P) residues, and the proteins are self-avoiding chains arranged on a simple cubic lattice. The H-H, H-P, and P-P pairwise interaction energies are defined as $\varepsilon_{HH} = -1, \varepsilon_{HP} = 0$, and $\varepsilon_{PP} = 0$, and only nearest-neighbor interactions are considered. The total energy of a chain in conformation s is thus given by

$$E(s) = n_{HH}\varepsilon_{HH}, \tag{1}$$

where n_{HH} is the number of H-H contact pairs non-adjacent on the chain. Under this energy function, a folded HP chain usually consists of a hydrophobic core and a polar shell. It mimics real proteins in native states with hydrophobic residues hidden from the solvent (water). Computational expense in evaluating energy of the HP model has been largely simplified. Nevertheless, the rough energy landscape of real proteins is somewhat retained, so it is still an ideal model to test and drive the development of efficient algorithms which target at finding global minima in a highly rough energy landscape.

Stochastic algorithms that simulate from the HP model generally fall into two categories [5]: chain-growth algorithms [10–12] and Markov chain Monte Carlo (MCMC) [6–8,13,14]. The idea of chain-growth algorithms is to construct a protein folding structure from the first residue by adding one residue at a time. Empty sites in neighbor to the leading residue carry different probabilities to accept a new residue, depending on the energy change induced by the new residue. Successful growth of the chain to the targeted length would be a valid sample. If a dead-end is reached, the growth process starts over from the first residue. Although the method may be efficient for short HP models, the idea is hard to transfer to the folding of real proteins. In comparison, an MCMC method starts from a whole chain, and the conformation evolves according to a specific move set. It is also referred to as a "blind search" method, because it does not require *a priori* knowledge on the protein native state, while chain-growth algorithms are usually designed to preferably form dense conformations. A popular choice of the move set is the *pull move*, which is local, reversible, and complete [15]. The pull move folds a chain locally, creates "humps" on the chain, and gradually forms dense structures. Details on how the pull move operates can be found in [6]. One drawback of the pull move is that when the chain is in highly dense conformations (quasi-folded states), only residues on the shell can perform valid pull moves, while residues in the core are relatively hard to change, and thus, the efficiency in proposing new conformations is largely reduced. The bond-rebridging move [8, 16] has been proposed to overcome this difficulty, by breaking and reconnecting bonds in the hydrophobic core. In this work, we combine the pull move and the bond-rebridging move as the move set for a lattice protein chain.

However, a straight-forward Metropolis algorithm on the HP model usually suffers from being trapped at local minima, because proposals to high-energy barriers surrounding a local minimum have an exponentially small acceptance probability. Various strategies have been proposed to alleviate this problem, including multicanonical sampling [17,18], entropic sampling [19], simulated tempering [20], evolutionary Monte Carlo [21], the equi-energy sampler [6], replica exchange Monte Carlo [14, 22], the Wang-Landau (WL) algorithm [7, 8], and gradient-directed Monte Carlo [23].

In reality, due to the continuum nature of folding of real proteins, there is only one global minimum (the native state). However, under moderate interruption of environmental conditions (e.g., temperature, pH value, ion concentrations), proteins may have a good chance to misfold into metastable states with energy close to the global minimum and lose their desired function, which is the cause of many diseases. The native state is typically only 5-10 kcal/mol lower in energy than a misfolded state [24]. Therefore, knowledge about those metastable states is very important for understanding the protein thermodynamics and further development of medications to help correct misfolded proteins. In this work, we develop an MCMC method to find multiple global minima and a large number of local minima with energy close to the global minima of the HP model. With the use of a new distance measure, minima found by our method all have distinct hydrophobic cores and represent conformations with non-trivial structural differences. This clearly distinguishes our method from existing MCMC methods which focus on thermodynamic estimation and often provide only a few minima, such as those reviewed above.

2 Methodology

We apply a recently developed algorithm, the multidomain sampler (MDS) [25], to the HP model. The MDS carries on and develop the idea of the WL algorithm [26, 27]. It not only performs a random walk over energy space, but also over different basins of attraction of local minima. The effectiveness of this algorithm in statistical inference and in constructing Ising energy landscapes has been demonstrated in our previous work [25, 28]. In this paper, we show that the MDS with suitable modifications can also serve as a powerful method for finding multiple global/local minima, because it avoids redundant visits to basins with adequate samples and pushes the sampler to less-explored portions of the state space. After a description of the MDS in the context of the HP model, we develop new strategies to achieve an efficient search for a large number of minimum-energy folding conformations.

2.1 Multidomain sampler

For a given energy function, the MDS can be used as an algorithm to search for its K lowest energy minima. We use dynamically updated information of K local minima, denoted as

$v_1, \ldots, v_K$, to partition the state space into $K+1$ domains, $D_1, \ldots, D_K$, and their complement D_0. For any state s, if it finds v_k by steepest descent, we say $s \in D_k$. In other words, D_k is the basin of attraction of v_k. To enable steepest descent in a discrete model like the HP model, neighbors of a given state s need to be defined. For the HP model, the set of neighbors of s is defined as all the conformations that can be accessed by a single pull move from the conformation s, as pull moves only locally mutate the original state. Then a steepest descent algorithm is implemented by recursive application of single pull moves that give the maximum energy decrease in each step until a local minimum is reached. When a new minimum with energy lower than the maximum energy of the K stored local minima is found, the new minimum replaces the highest-energy minimum in the original set. Hence, the set $\mathcal{V} = \{v_k : k = 1, \ldots, K\}$ always keeps the K minima with the lowest energy so far. When the steepest descent finds a local minimum with energy higher than those in $\mathcal{V}$, the conformation s is assigned to D_0. On the other hand, similar to the WL algorithm, the energy space is also partitioned by a ladder of energies, $u_1 < ... < u_L < u_{L+1} = \infty$. Usually the global minimum of a given chain is unknown, so the ladder is dynamically updated during the simulation process to ensure the lowest minimum found so far is in $[u_1, u_2)$. The objective for this double-partitioning design is to drive the sampler to perform a random walk over all non-empty subregions,

$$D_{kj} = \{s \in D_k : E(s) \in [u_j, u_{j+1})\},$$

$k = 0, \ldots, K$, $j = 1, \ldots, L$, and also to encourage the sampler to explore thoroughly the basins of newly found low-energy minima.

The mechanism to generate a random walk over all D_{kj} is similar to the generalized Wang-Landau (GWL) algorithm [29, 30], in which each energy interval may contain multiple energy levels. Let $\theta_{kj} \propto \sum_{s \in D_{kj}} e^{-\beta E(s)}$ denote the statistical weight (unnormalized) of D_{kj} in the Boltzmann distribution at temperature $T = 1/\beta$. A flat histogram can be generated if the probability of visiting a state $s \in D_{kj}$ is proportional to $e^{-\beta E(s)}/\theta_{kj}$. The weights θ_{kj} can be estimated by a WL-type iterative algorithm. At the t^{th} iteration, θ_{kj} is estimated by $\theta_{kj}^{(t)}$ ($\theta_{kj}^{(1)}$ is set to 1 for all k and j). If the state at this iteration is $x_t \in D_{kj}$ and a new state $y \in D_{\ell i}$ is proposed, the Metropolis ratio from x_t to y is

$$r(x_t \to y) = \min \left\{ 1, e^{\beta[E(x_t) - E(y)]} \frac{\theta_{kj}^{(t)}}{\theta_{\ell i}^{(t)}} \frac{P(y \to x_t)}{P(x_t \to y)} \right\}, \tag{2}$$

where $P(y \to x_t)$ is the proposal probability from y to x_t and $P(x_t \to y)$ the proposal probability from x_t to y. In addition to the pull move, our proposals for new conformations also include the bond-rebridging move [8, 16], which is a good complement as it is more efficient than the pull move for updating dense conformations. The probability to propose a pull move, denoted by P_{pm}, is fixed to 0.9 throughout our simulation. Denote by x_{t+1} the updated state according

to the above Metropolis ratio. Then

$$\ln \theta_{kj}^{(t+1)} = \ln \theta_{kj}^{(t)} + \mathbf{1}(x_{t+1} \in D_{kj}) \ln f \tag{3}$$

is used to update the estimation of θ_{kj} for all k, j, where $f > 1$ ($\ln f > 0$) is the modification factor and $\mathbf{1}(x_{t+1} \in D_{kj}) = 1$ if $x_{t+1} \in D_{kj}$ and 0 otherwise. It is easy to see that the Metropolis ratio increases with $\theta_{kj}^{(t)}/\theta_{\ell i}^{(t)}$. Thus, if the basin D_k has already been visited frequently, a proposed y to another basin will have a higher acceptance rate. Unlike those works that use the WL algorithm to determine the density of states of the HP model [8, 31, 32], we do not reduce f to achieve convergence in thermodynamic estimation. Since our goal here is to find as many minima as possible, while sampling is less of a concern, we keep $f \equiv e$ ($\ln f \equiv 1$) throughout the simulation to ensure fast growth in $\theta_{kj}^{(t)}$ and therefore an fast exploration over many basins. Under this setting, the log-weight $\ln \theta_{kj}^{(t)}$ simply records the number of visits to D_{kj} until the current iteration. This makes it convenient to update these weights when the set of stored minima $\mathcal{V}$ is updated. Here, note that steepest descent with the pull move is used to determine the basin of the proposed conformation y. Suppose it finds the local minimum $v(y)$. If $v(y)$ replaces an existing minimum v_m in $\mathcal{V}$, then we add the current log-weights of the basin D_m, i.e., $\ln \theta_{mj}^{(t)}$ for all j, to the log-weights of D_0, because D_m now becomes a part of D_0, and reset $\ln \theta_{mj}^{(t)} = 0$ as the initial weights for the basin of the new minimum $v(y)$.

In other applications when sampling is the primary goal, we can first run the MDS with $f \equiv e$ for a while, updating dynamically the set of minima $\mathcal{V}$, and then gradually reduce f with $\mathcal{V}$ fixed to achieve convergence in sampling. This strategy has been adopted in our previous work [25] on structural sampling of Bayesian networks, where we have obtained reliable and accurate estimation of various statistics of interest. This shows that minima found by fixing $f \equiv e$ are often good representatives for the low-energy portion of the overall landscape.

2.2 Visit-enhancement factor

In principle, large K is preferred, because more local minima can be recorded, which would be helpful to guide the random walk. However, the conformation space of a HP lattice protein, as well as the total number of local minima, grows exponentially with the chain length [33]. Clearly, K cannot scale exponentially due to limited computing resources. For a long chain (e.g., 100 residues), $\bigcup_{k=1}^{K} D_k$ usually represents only a small portion of the conformation space, and its complement D_0 takes up the major part. Various portions in D_0 may be connected to unknown global minima or low-energy local minima, and thus, adequate transitions between D_0 and other basins are important. In fact, considering the huge number of conformations in D_0, it is unfair to treat it as an ordinary domain for building a flat histogram. Consequently, we increase the chance to visit D_0 by the use of a visit-enhancement factor A.

Suppose the desired frequency of visiting each D_{kj} ($k \geq 1$) is ψ and that of each D_{0j} is $A\psi$

($A > 1$). To achieve the desired frequencies, one can modify (3) to

$$\ln \theta_{kj}^{(t+1)} = \ln \theta_{kj}^{(t)} + [\mathbf{1}(x_{t+1} \in D_{kj}) - \psi] \ln f, \text{ for } k \geq 1, \quad (4)$$

$$\ln \theta_{0j}^{(t+1)} = \ln \theta_{0j}^{(t)} + [\mathbf{1}(x_{t+1} \in D_{0j}) - A\psi] \ln f, \quad (5)$$

as shown by Liang *et al.* [29]. Since adding a constant $\psi \ln f$ to the right-hand sides of the above recursions does not change the Metropolis ratio (2) for the $(t+1)^{\text{th}}$ iteration, we may still use (3) for $k \geq 1$ but use

$$\ln \theta_{0j}^{(t+1)} = \ln \theta_{0j}^{(t)} + [\mathbf{1}(x_{t+1} \in D_{0j}) - (A-1)\psi] \ln f \quad (6)$$

for D_0. The exact value of ψ is determined by the numbers of non-empty subregions in D_k ($k \geq 1$) and D_0, which may change when the set of stored minima $\mathcal{V}$ is updated. Since the visit-enhancement factor A is essentially a tuning parameter to tune the search from depth-first to breadth-first as A increases, we simply fix ψ to $1/(A+K)$, which is the exact value when $L = 1$. The weight $\theta_{0j}^{(t+1)}$ decreases if the domain D_0 is not visited. The larger the A is, the faster $\theta_{0j}^{(t+1)}$ decreases. Hence, acceptance of proposals to D_0 becomes easier over time. The domain D_0 contains high-energy "open" protein conformations, and the chain needs to first unfold from a "close" form and then may be able to fold into a different conformation. Therefore, D_0 serves as a pathway for the sampler to jump between different local basins and visit unexplored part of the space.

Now we give a complete outline of the MDS for the HP model. Define the domain partition index of a state s by $I(s) = k$ if $s \in D_k$ and the energy partition index $J(s) = j$ if $E(s) \in [u_j, u_{j+1})$. Given x_t and $\{\theta_{kj}^{(t)}\}$, one iteration of our algorithm is composed of the following steps.

1. Propose a conformation y with probability P_{pm} by a pull move or with probability $1 - P_{pm}$ by a bond-rebridging move.

2. Perform a steepest descent search to find $v(y)$. Update the set of local minima $\mathcal{V}$ and the energy ladder if needed. Then determine the domain partition index $I(y)$ and the energy partition index $J(y)$.

3. Accept or reject y via the Metropolis ratio (2) with $k = I(x_t)$, $j = J(x_t)$, $\ell = I(y)$ and $i = J(y)$ to obtain the updated conformation x_{t+1}. Use (3) and (6) to update the weights $\theta_{kj}^{(t+1)}$ for all k, j.

2.3 H-residue subchain distance

In step 2 of the algorithm outline, when a new conformation y is proposed and $v(y)$ is found, a one-by-one comparison between $v(y)$ and $\mathcal{V}$ is performed to determine whether $v(y)$ matches any

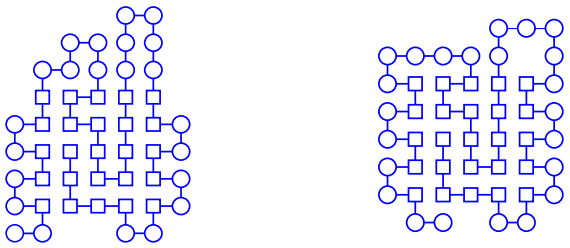


Figure 1: Different folding structures in a strict distance definition but having zero HSC distance. H and P residues are represented by □ and ◯, respectively.

of the K stored minima. The comparison is achieved by computing a distance metric $d(s^{(1)}, s^{(2)})$ between two conformations $s^{(1)}$ and $s^{(2)}$. If $d(s^{(1)}, s^{(2)}) = 0$, then we say $s^{(1)}$ matches $s^{(2)}$. Obviously, identical conformations should match each other under any distance metric. For a specific distance, if $d(s^{(1)}, s^{(2)}) = 0$ only when $s^{(1)}$ and $s^{(2)}$ are identical, then we say it is a *strict* distance definition; otherwise it is a *loose* distance definition. An example of a strict distance definition is described in [4]. Each residue is coded using the direction of its following bond with respect to its previous bond, which is either 0 (collinear), +1 (right turn), or −1 (left turn). Therefore, each conformation can be described by a vector, and the distance between two conformations is defined by the L_1 norm (sum of the absolute value of each component) of the difference between their corresponding vectors. However, in the HP model, only the arrangement of H residues determines the total energy of the chain, and P residues are "dummy". A strict distance metric differentiates similar conformations like the two shown in Figure 1, which have the same H-residue arrangement but differ in P-residues. In our algorithm, since only a limited number K of local minima can be stored to guide the search, it is a good idea to make them as diverse as possible, so that the K local minima can represent a larger portion of the space as a whole. We thus introduce a new distance metric that only measures the difference in the H-residue subchain (HSC) when comparing two conformations. Consequently, the basins of minima having the same arrangement of the HSC are merged into one domain in our algorithm.

Assuming the length of a chain is l, which contains n H-residues, the HSC of the chain $s = (s_1, s_2, \ldots, s_l)$ is defined as $s_H = (s_{1,H}, s_{2,H}, \ldots, s_{n,H})$, where $s_{j,H}$ is the jth H-residue. For two conformations of the chain, $s^{(1)}$ and $s^{(2)}$, their HSC distance is

$$d(s^{(1)}, s^{(2)}) = \sum_{j=2}^{n-1} d_j(s_H^{(1)}, s_H^{(2)}), \tag{7}$$

with $s_H^{(k)}$ $(k = 1, 2)$ being the conformation of the HSC. The definition of $d_j(s_H^{(1)}, s_H^{(2)})$ is described below and illustrated in Figure 2(a). Let $\vec{w}_j^{(k)} = s_{j,H}^{(k)} - s_{j-1,H}^{(k)}$, where $s_{j,H}^{(k)}$ is the vector of the coordinates of the jth H-residue in conformation $s^{(k)}$. When calculating $d_j(s_H^{(1)}, s_H^{(2)})$, $s^{(2)}$ is

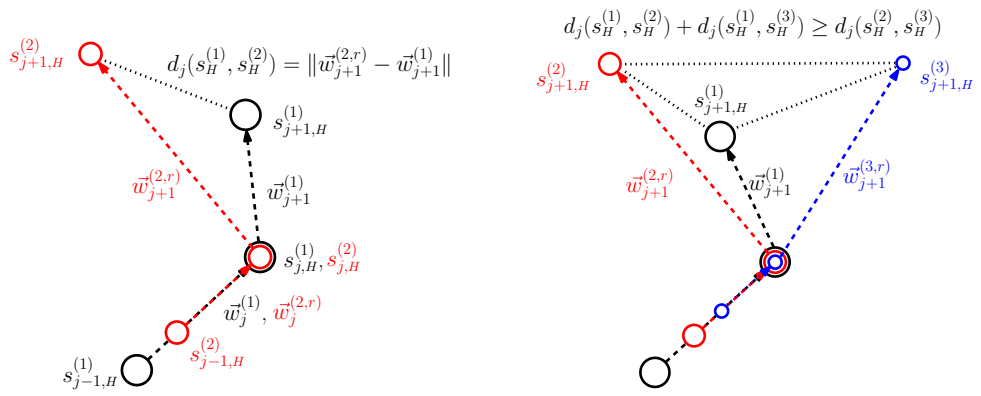


Figure 2: (Color online) (a) Schematics of the HSC distance definition. HSC in two conformations is represented by circles of different sizes and colors. Conformation $s^{(2)}$ in red has been rotated so that $\vec{w}^{(2,r)}_j$ is parallel to $\vec{w}^{(1)}_j$. Dashed lines indicate that the H-residues are not necessarily adjacent to each other. The Euclidean norm of the dotted line segment gives $d_j(s^{(1)}_H, s^{(2)}_H)$. (b) Triangle inequality in the HSC distance definition. Conformations in red and blue have been rotated. Three conformations involved are distinguished by circles with different sizes.

rotated so that $\vec{w}^{(2)}_j$ is parallel to $\vec{w}^{(1)}_j$ [see Figure 2(a)]. Denote $\vec{w}^{(k)}_j$ and $\vec{w}^{(k)}_{j+1}$ after rotation by $\vec{w}^{(k,r)}_j$ and $\vec{w}^{(k,r)}_{j+1}$. Then, $d_j(s^{(1)}_H, s^{(2)}_H) = \|\vec{w}^{(2,r)}_{j+1} - \vec{w}^{(1)}_{j+1}\|$, where $\|\cdot\|$ is the Euclidean norm of a 2D vector. The HSC distance satisfies the general requirements of a distance definition. First, it is always non-negative. Second, if instead, $s^{(1)}$ is rotated to align $\vec{w}^{(1)}_j$ to $\vec{w}^{(2)}_j$, it is easy to show that

$$d_j(s^{(2)}_H, s^{(1)}_H) = \|\vec{w}^{(1,r)}_{j+1} - \vec{w}^{(2)}_{j+1}\| = \|\vec{w}^{(2,r)}_{j+1} - \vec{w}^{(1)}_{j+1}\| = d_j(s^{(1)}_H, s^{(2)}_H). \tag{8}$$

The distance is symmetric at each residue and thus symmetric for the whole chain. Third, when comparing three conformations $s^{(1)}$, $s^{(2)}$ and $s^{(3)}$, $s^{(2)}$ and $s^{(3)}$ are rotated to align $\vec{w}^{(2)}_j$ and $\vec{w}^{(3)}_j$ to $\vec{w}^{(1)}_j$. Note that the inequality

$$\begin{aligned} d_j(s^{(1)}_H, s^{(2)}_H) + d_j(s^{(1)}_H, s^{(3)}_H) &= \|\vec{w}^{(2,r)}_{j+1} - \vec{w}^{(1)}_{j+1}\| + \|\vec{w}^{(3,r)}_{j+1} - \vec{w}^{(1)}_{j+1}\| \\ &\geq \|\vec{w}^{(2,r)}_{j+1} - \vec{w}^{(3,r)}_{j+1}\| = d_j(s^{(2)}_H, s^{(3)}_H) \end{aligned} \tag{9}$$

holds at each residue, which implies that the distance definition satisfies triangle inequality. See Figure 2(b) for an illustration. The HSC distance definition is also extensible to a 3D chain, with an additional rotation operation. Similarly, we first align $\vec{w}^{(2)}_j$ to $\vec{w}^{(1)}_j$. Then rotate $s^{(2)}$ around $\vec{w}^{(2,r)}_j$ such that $\|\vec{w}^{(2,r)}_{j+1} - \vec{w}^{(1)}_{j+1}\|$ is minimized, which occurs when $\vec{w}^{(2,r)}_{j+1}$ is in the plane spanned by $\vec{w}^{(1)}_j$ and $\vec{w}^{(1)}_{j+1}$. After the rotation, the same HSC distance definition can be used.

To demonstrate the efficiency of our algorithm, we compare it with a two-step search algorithm, which first uses the GWL algorithm [29] with $f \equiv e$ to sample conformations and then

find the associated minimum of each conformation by steepest descent. This two-step approach does not utilize the partition by the basins of attraction of local minima, but other parameter settings are exactly identical to those used in the MDS. The probability to accept a new conformation y given the current x_t is

$$r(x_t \to y) = \min \left\{ 1, e^{\beta[E(x_t)-E(y)]} \frac{\theta^{(t)}_{J(x_t)}}{\theta^{(t)}_{J(y)}} \frac{P(y \to x_t)}{P(x_t \to y)} \right\}, \tag{10}$$

where $J(\cdot)$ gives the energy partition index and $\theta^{(t)}_j$ is the current estimate of the weight of the jth energy interval, $j = 1, \ldots, L$. These weights are then updated by

$$\ln \theta^{(t+1)}_j = \ln \theta^{(t)}_j + \mathbf{1}(J(x_{t+1}) = j) \tag{11}$$

as $\ln f \equiv 1$, where x_{t+1} is the updated conformation according to the above Metropolis ratio.

3 Results

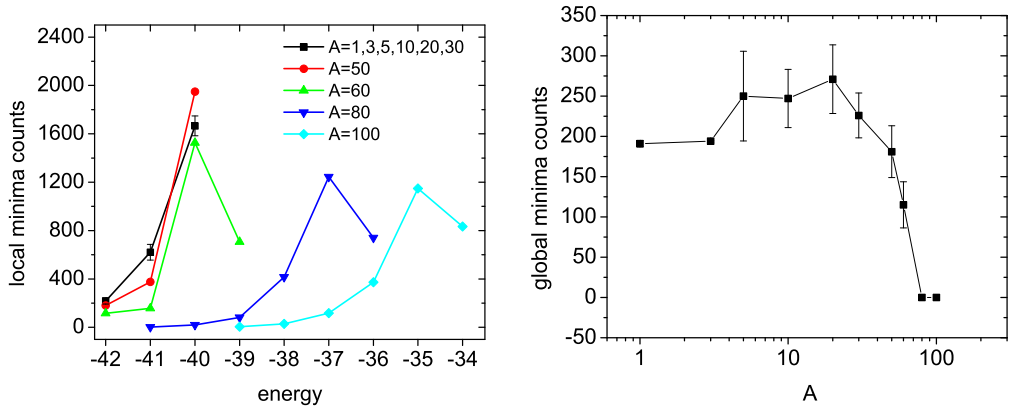
The proposed algorithm was applied to three benchmark 2D protein sequences: seq48, seq64, and seq100b [6]. See Table 1 for their residue sequences. The number of MC steps was 5×10^6 for each individual run. The energy ladder generally included 10 evenly distributed intervals and we chose $K = 500$. The initial conformations were straight chains for all sequences in all runs.

Table 1: Residual sequences of three benchmark lattice proteins

name	residue sequence
seq48	PPHPPHHPPHHPPPPPHHHHHHHH HHPPPPPPHHPPHHPPHPPHHHHH
seq64	HHHHHHHHHHHHPHPHPPHHPPHH PPHPPHHPPHHPPHPPHHPPHHPPH PHPHHHHHHHHHHHH
seq100b	PPPHHPPHHHHPPHHHPHHPHHPHH HHPPPPPPPPHHHHHHPPHHHHHHP PPPPPPPPHPHHPHHHHHHHHHHP PHHHPHHPHPPHPHHHPPPPPPHHH

3.1 Performance evaluation

We first examined the effect of the visit-enhancement factor A on finding low-energy minima. Figure 3(a) shows the energy distributions of the K lowest-energy minima of seq64 found by the MDS across a wide range of values of A, from 1 to 100. The counts reported are the sums of



(a) (b)

Figure 3: (Color online) (a) Energy distributions of the K lowest minima found for different visit-enhancement factor A. The counts are the sums of five individual runs. The curves for $A \leq 30$ are summarized into a single curve of the mean values, due to their similarity, with standard deviations indicated by error bars. (b) Number of global minima found for different A, where the error bar of a data point is the estimated standard deviation of the total count. A missing error bar implies that the standard deviation is very small.

five individual runs. Figure 3(b) shows the total number of global minima (energy $E_g = -42$) found over five individual runs with different A. The sampler showed optimal and comparable performance in finding global and low-energy minima for A between 5 and 30, and it found significantly fewer low-energy minima for $A \geq 50$. In the extreme scenario, the sampler failed to find any global minima when $A \geq 80$. This is expected as the sampler tends to spend too much time in the complement domain D_0 when A is too big, and thus, may not explore thoroughly those basins of low-energy minima. In practice, we found that the optimal value of A is not sensitive to the length of a chain and thus fixed $A = 20$ for simulating all the sequences in this study.

Table 2 gives a summary of the results for all three sequences. The energy of the lowest minimum found in the literature [6–8, 13, 34] for each sequence is reported in the table. If our algorithm finds a minimum with that energy, we call it a global minimum. The total numbers of global minima found by the MDS over five individual runs were 70, 271, and 12 for seq48, seq64, and seq100b, respectively. This shows that the MDS indeed is able to identify multiple, sometimes many, global minima. Note that our algorithm stored 500 local minima (including global minima) in every run. Statistics such as the average and the standard deviation of the energies of all 2500 minima over five runs are reported in the table. The average energies were often only one to two units higher than the global minima and the standard deviations were also small. This shows that our algorithm also detected a large number of local minima with energy very comparable to that of the global minima. On the other hand, the average pairwise HSC distance between the identified minima confirms that these folding conformations were

quite different in their hydrophobic cores. These results clearly demonstrate that the HP model has a large number of minimum-energy conformations, and thus, it is a great advantage for understanding the overall energy landscape to be able to find a substantial number of these minima with high diversity.

Table 2: Statistics of minima found by the MDS and the GWL

sequence	method	E_g^*	E_g	N_g	$\bar{E}$	SD	distance
seq48	MDS	-23	-23	70	-22.0	0.4	23.6
	GWL		-23	32	-21.2	0.4	24.5
seq64	MDS	-42	-42	271	-40.6	0.7	31.8
	GWL		-42	4	-38.2	0.7	40.1
seq100b	MDS	-50	-50	12	-47.7	1.0	43.7
	GWL		-48	0	-45.0	1.2	54.0

E_g^*: reported lowest minimum energy in the literature; E_g: global minimum energy found in this work; N_g: total number of global minima found; $\bar{E}$: average energy of all stored local minima; SD: standard deviation of the energies of all stored local minima; distance: average pairwise distance between stored local minima. Results from five runs are pooled together in the table.

To further benchmark the performance of the MDS, we applied the GWL-based two-step approach to the same set of sequences, with the same number of MC steps and the same energy ladders. A side-by-side comparison between the two methods is given in Table 2. It is noted that the MDS always found much more global minima than the GWL algorithm did, and the latter failed to find any global minima for seq100b. The average energy of identified local/global minima by our algorithm was lower than the GWL algorithm, which demonstrates the effectiveness of the MDS in folding lattice proteins. Because high-energy conformations are often more open and therefore dissimilar, the minima found by the GWL had a higher average pairwise distance. This comparison shows the usefulness of generating a random walk over different basins of attraction via the double-partitioning design in detecting low-energy minima.

3.2 Clustering global minima

As demonstrated by Table 2, the MDS can efficiently find multiple global minima and many low-energy local minima of a given protein chain. It is helpful to examine systematically the diversity among all the global minima found for a specific protein sequence. To this end, we used single-linkage hierarchical clustering to group global minima with the HSC distance as the measure of dissimilarity. Figure 4 and Figure 5 show example cluster trees for the three chains. The height of an internal node represents the HSC distance between the nearest neighbors in the two sub-clusters. Two major clusters are readily distinguishable in Figure 4(a) and (b), where the conformations in one cluster are the mirror images of those in the other cluster. In our HSC

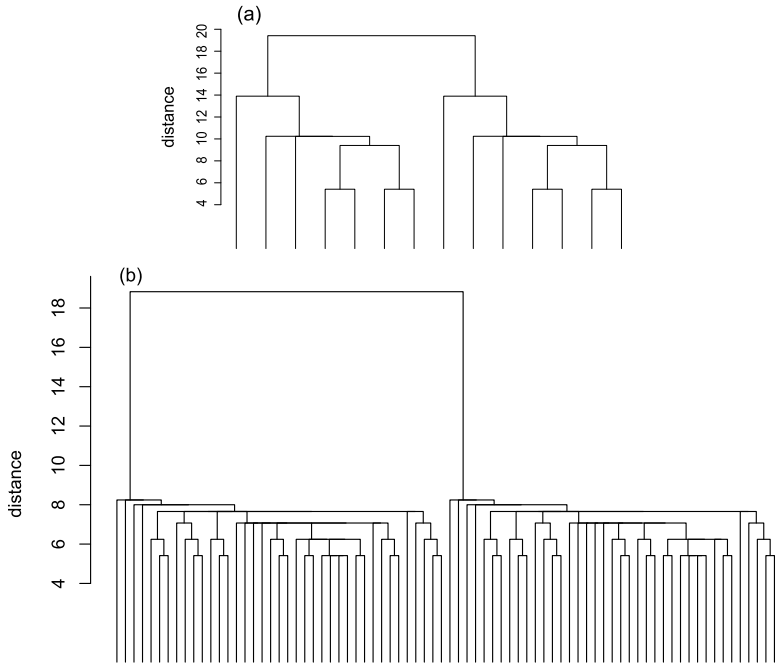


Figure 4: Hierarchical clustering of global minima found in a single run: (a) seq48, (b) seq64.

distance measure, a certain conformation and its mirror image are treated as different conformations. In fact, molecules may have different chemical properties from their mirror images, which is referred to as chirality, and thus it is biologically meaningful to treat a conformation and its mirror image differently.

For seq48, all the individual runs identified the same set of 14 global minima, with seven in each of the mirror image clusters [Figure 4(a)], even without visit enhancement for D_0 (i.e., $A = 1$). Therefore, the introduction of A seems less important for short chains. There were 78 global minima found for seq64 in the most successful run (finding the most global minima), with 39 in each of the two mirror image clusters [Figure 4(b)]. The hierarchical structures within the two clusters are essentially identical, up to a permutation of the horizontal placement of the global minima in one cluster. All the global minima found in other runs were subsets of the above 78 global minima, which suggests they may represent a complete set of the global minima for this sequence. For seq100b, the minima found were relatively diverse, but did not contain the corresponding mirror image cluster due to the large conformation space and a limited number of search steps. A protein in a compact folding state cannot easily evolve to its mirror image conformation unless the chain sufficiently unfolds. Therefore, a high percentage of minima that form mirror image pairs may be regarded as evidence of an efficient random walk in the conformation space. The match percentage is defined as $\frac{2k}{n_T}$, where k is the number of matched image mirror pairs and n_T is the total number of global minima found. For example, the match

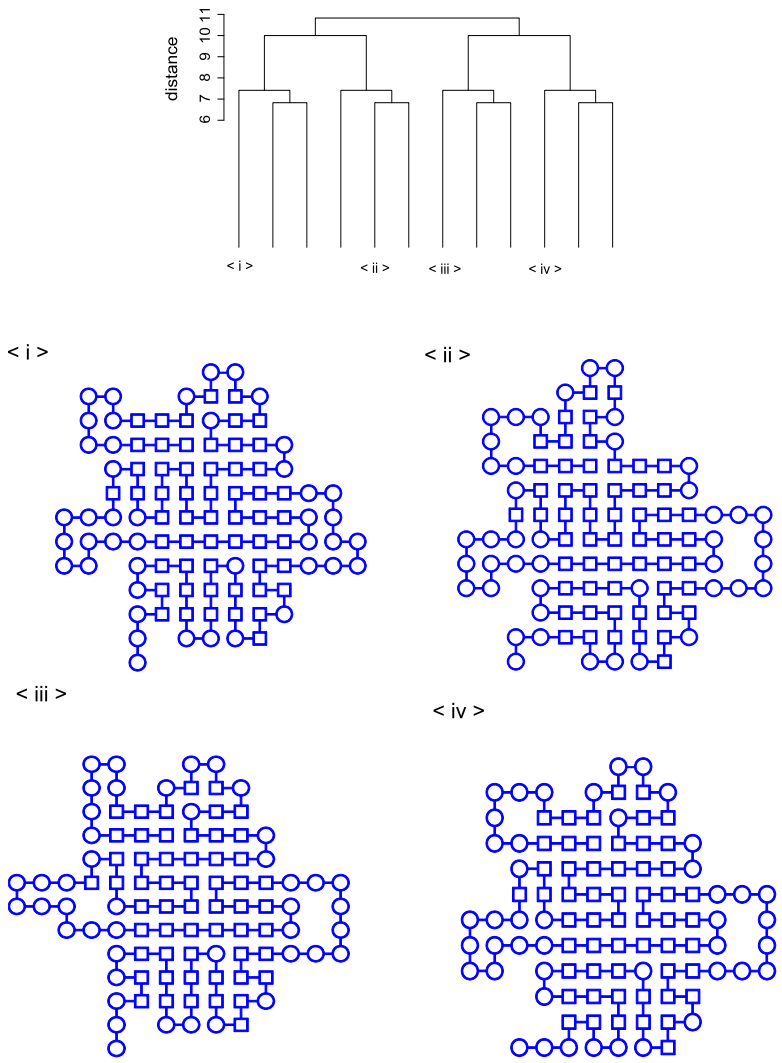


Figure 5: Hierarchical clustering of global minima of seq100b found in a single run and the folding conformations of four global minima, each from a cluster.

percentages of the global minima on the two cluster trees in Figure 4 are both 100%. The overall match percentages across five independent runs for seq48, seq64, and seq100b were 100%, 63%, and 0%, respectively. The low match quality of the global minima found for seq100b implies an insufficient random walk over the space. However, this sequence has been challenging for many algorithms. For example, the equi-energy sampler failed to find any global minima for this sequence [6]. Although the MDS could not move around the space so freely as to reach mirror image pairs, it indeed identified in a single run 12 distinct global minima, with pairwise HSC distances between 6 and 11 (Figure 5). If we use a distance cutoff around 9, these global minima can be grouped into four clusters according to the cluster tree. The conformations of four global minima, one from each cluster, are shown below the cluster tree in Figure 5. One

sees that these conformations all have quite different hydrophobic cores, and a direct proposal from one of them to another by the move set is almost impossible. Thus, the sampler must first climb up a high energy barrier by sufficiently unfolding a compact conformation and then move downhill to reach another global minimum. The double-partitioning design is the key to achieving such free moves between different energy levels and across different basins of attraction. To the best of our knowledge, this result is the first explicit demonstration of finding multiple diverse global minima for this sequence in a single MCMC simulation. Lastly, to facilitate future analysis, all the distinct global minima found for the three sequences in this study are provided at `www.stat.ucla.edu/~zhou/MDSHP/` and in the supplemental material.

4 Discussion

In this work, we have demonstrated the power of the MDS in searching for multiple distinct global minima of various benchmark HP lattice protein sequences. By extending the idea in the WL algorithm of partitioning energy space, the MDS introduces additional partitioning of the conformation space into the basins of attraction of local minima. This double-partitioning design significantly increases the efficiency of the sampler in exploring unvisited part of the space by avoiding redundant sampling from the same domain. A visit-enhancement factor is introduced to facilitate jumps between $\{D_k : k = 1, \ldots, K\}$ and their complement D_0, as the latter usually contains important pathways to unexplored portion of the space.

One unique aspect of the MDS is that it utilizes information on local basin structures. In addition to the advantage in finding minima reported in this paper, this information is also useful for estimating barriers between different basins and constructing disconnectivity graphs [35, 36] as demonstrated in [28] on spin glasses. Practically, when a protein misfolds into a metastable state with a high barrier to access other domains, the protein will be trapped in this local basin and therefore lose its desired function. Estimating cross-domain barriers helps to identify those deep metastable states, and may provide crucial fundamental understandings of protein behavior in drug design. One of our future directions is to develop efficient methods to construct energy barriers based on conformations simulated from the MDS.

Moreover, in the MDS algorithm, we use information from K stored minima to define domains and to guide search for global minima. We would like to maximize the volume represented by the K domains in the conformation space, while keeping the number K relatively small so that each domain can have sufficient samples given an upper limit on the total number of conformations in a simulation. Proper schemes to merge similar and locally connected domains are therefore highly desired. To this end, we have defined a new HSC distance for the HP model, which implicitly merges those basins whose local minima only differ in the arrangement of P-residues. Development of more general schemes to dynamically merge basins/domains separated by low barriers [37] is another interesting topic for future investigation.

It is noted that modifications of the WL algorithm have been proposed in the literature to improve its sampling efficiency. For example, multiple runs of the WL algorithm have been used to generate multiple random walks, each restricted to an energy window that slightly overlaps adjacent ones. Cunha-Netto *et al.* developed an adaptive window approach to alleviate the border effect when multiple energy windows are used for a large system [38]. This approach is clearly different from the MDS, where a random walk is simulated over the entire energy range. The additional domain partitioning is on the conformation space, not the energy space. However, our algorithm may also benefit by similar ideas as the adaptive window approach when sample from or optimize over large systems.

Acknowledgment

This work was supported by an NSF award DMS-1055286 to Q.Z.

References

[1] P.L. Freddolino, C.B. Harrison, and Y. Liu, K. Schulten, Nat. Phys. **6**, 751 (2010).

[2] J. Shimada and E.I. Shakhnovich, Proc. Nalt. Acad. Sci. **99**, 11175 (2011).

[3] J.H. Meinke and U.H. Hansmann, J. Comput. Chem. **30**, 1642 (2009).

[4] K.F. Lau and K.A. Dill, Macromolecules **22**, 3987 (1989).

[5] M. Bachmann and W. Janke, J. Chem. Phys. **120**, 6770 (2004).

[6] S.C. Kou, J. Oh, and W.H. Wong, J. Chem. Phys. **124**, 244903 (2006).

[7] T. Wüst and D.P. Landau, Comput. Phys. Commun. **179**, 124 (2008).

[8] T. Wüst and D.P. Landau, Phys. Rev. Lett. **102**, 178101 (2009).

[9] J. Liu, G. Li, and J. Yu, Phys. Rev. E **84**, 031934 (2011).

[10] J.L. Zhang and J.S. Liu, J. Chem. Phys. **117**, 3492 (2002).

[11] H.P. Hsu, V. Mehra, W. Nadler, and P Grassberger, Phys. Rev. E **68**, 021113 (2003).

[12] W. Huang, Z. Lu, and H. Shi, Phys. Rev. E **72**, 016704 (2005).

[13] J. Zhang, S.C. Kou, and J.S. Liu, J. Chem. Phys. **126**, 225101 (2007).

[14] C. Thachuk, A. Shmygelska, and H.H. Hoos, BMC Bioinformatics **8**, 342 (2007).

[15] N. Lesh, M. Mitzenmacher, and S. Whitesides, Proc. RECOMB (2003).

[16] J.M. Deutsch, J. Chem. Phys. **106**, 8849 (1997).

[17] B.A. Berg and T. Neuhaus, Phys. Rev. Lett. **68**, 9 (1992).

[18] H. Arkin, J. Stat. Phys. **139**, 326 (2010).

[19] J. Lee, Phys. Rev. Lett. **71**, 211 (1993).

[20] A.P. Lyubartsev, A.A. Martsinovski, S.V. Shevkunov, and P.N. Vorontsov-Velyaminov., J. Chem. Phys. **96**, 1776 (1992).

[21] F. Liang and W.H. Wong, J. Chem. Phys. **115**, 3374 (2001).

[22] D. Gront, A. Kolinski, and J. Skolnick, J. Chem. Phys. **115**, 1569 (2001).

[23] X. Hu, D.N. Beratan, and W. Yang, J. Chem. Phys. **131**, 154117 (2009).

[24] K.A. Dill, S.B. Ozkan, M.S. Shell, and T.R. Weikl, Annu. Rev. Biophys. **37**, 289 (2008).

[25] Q. Zhou, J. Am. Stat. Assoc. **106**, 1317 (2011).

[26] F. Wang and D.P. Landau, Phys. Rev. Lett. **86**, 2050 (2001).

[27] F. Wang and D.P. Landau, Phys. Rev. E **64**, 056101 (2001).

[28] Q. Zhou, Phys. Rev. Lett. **106**, 180602 (2011).

[29] F. Liang, C. Liu, and R.J. Carroll, J. Am. Stat. Assoc. **102**, 305 (2007).

[30] Y.F. Atchade and J.S. Liu, Stat. Sinica **20**, 209 (2010).

[31] A.D. Swetnam and M.P. Allen, J. Comp. Chem. **32**, 816 (2011).

[32] Y.W. Li, T. Wüst, and D.P. Landau, Comput. Phys. Commun. **182**, 1896 (2011).

[33] R. Unger and J. Moult, J. Mol. Biol. **231**, 75 (1993).

[34] B. Chen and J. Hu, IEEJ Trans. Electrical and Electronic Eng. **5**, 459 (2010).

[35] O.M. Becker and M. Karplus, J. Chem. Phys. **106**, 1495 (1997).

[36] D.J. Wales, M.A. Miller, and T.R. Walsh, Nature **394**, 758 (1998).

[37] B. Strodel and D.J. Wales, Chem. Phys. Lett. **466**, 105 (2008).

[38] A.G. Cunha-Netto, A.A. Caparica, S.-H. Tsai, R. Dickman and D.P. Landau, Phys. Rev. E **78**, 055701(R) (2008).