PROTEINS: Structure, Function, and Genetics 48:192–201 (2002)

Local Energy Landscape Flattening: Parallel Hyperbolic Monte Carlo Sampling of Protein Folding

Yang Zhang, Daisuke Kihara, and Jeffrey Skolnick*
Laboratory of Computational Genomics, Donald Danforth Plant Science Center, St. Louis, Missouri

ABSTRACT **Among the major difficulties in protein structure prediction is the roughness of the energy landscape that must be searched for the global energy minimum. To address this issue, we have developed a novel Monte Carlo algorithm called parallel hyperbolic sampling (PHS) that logarithmically flattens local high-energy barriers and, therefore, allows the simulation to tunnel more efficiently through energetically inaccessible regions to low-energy valleys. Here, we show the utility of this approach by applying it to the SICHO (SIde-CHain-Only) protein model. For the same CPU time, the parallel hyperbolic sampling method can identify much lower energy states and explore a larger region phase space than the commonly used replica sampling (RS) Monte Carlo method. By clustering the simulated structures obtained in the PHS implementation of the SICHO model, we can successfully predict, among a representative benchmark 65 proteins set, 50 cases in which one of the top 5 clusters have a root-mean-square deviation (RMSD) from the native structure below 6.5 Å. Compared with our previous calculations that used RS as the conformational search procedure, the number of successful predictions increased by four and the CPU cost is reduced. By comparing the structure clusters produced by both PHS and RS, we find a strong correlation between the quality of predicted structures and the minimum relative RMSD (mrRMSD) of structures clusters identified by using different search engines. This mrRMSD correlation may be useful in blind prediction as an indicator of the likelihood of successful folds. Proteins 2002;48:192–201.**

INTRODUCTION

One of the key problems in the prediction of a protein's structure from its amino acid sequence is the development of a powerful optimization method that can find within a feasible amount of computer time the minimum energy structure; this corresponds to the native state according to the thermodynamic hypothesis of Anfinsen.[1] In principle, this global minimum energy state could be found from Metropolis Monte Carlo (MC) simulations where the probability of finding low-energy structures is exponentially enhanced compared with a random walk. However, because the energy landscape of the real protein sequences is characterized by numerous local minima separated by energy barriers, at low temperatures the Metropolis Monte Carlo scheme often gets trapped in these local minima. Thus, in practice, only small parts of the entire phase space are explored, thereby rendering the traditional Monte Carlo method impractical for protein structure prediction.[2–4]

One of the most efficient techniques in literature that is designed to overcome this local minima trapping is the replica sampling (RS) algorithm,[5,6] in which the simulations of several replicas are implemented at different temperatures. By exchanging the states at different temperatures, the higher-temperature process can help the lower-temperature structures cross the energy barriers between different basins and thereby achieve ergodicity. In a recent work,[7] we applied the RS technique to the simulation of a benchmark set of 65 proteins and used a reduced protein model, the SICHO (SIde-CHain Only) lattice model[8] to represent the protein. By clustering the produced structures, we have successfully predicted 46 proteins where one of the top five clusters has a root-mean-square deviation (RMSD) from the native state below 6.5 Å. However, although most of such cases could be found in one trajectory, to get all 46 cases, we found we needed to run at least 50 trajectories,† each costing about 72 h of CPU time on a 750-MHz Pentium III processor for an average length sequence (~100 residues). Obviously, this is very computationally demanding, and algorithms that reduce the requisite CPU time are needed.

Here, we propose a parallel hyperbolic Monte Carlo algorithm that can speed up the thermalization of the protein-folding process by flattening the local high-energy barriers found on the rough energy landscape. Thus, simulations at different temperatures are implemented on a dynamic, relatively smooth landscape. We apply this

Grant sponsor: National Institutes of Health; Grant number: GM-37408.

*Correspondence to: Jeffrey Skolnick, Laboratory of Computational Genomics, Donald Danforth Plant Science Center, 975 North Warson Road, Creve Coeur, MO 63132. E-mail: skolnick@danforthcenter.org

†We have run the clustering program based on 1 RS trajectory that is randomly chosen from the 50 trajectories produced in Ref. 7. On average, we can have 41.9 cases in which the best cluster has RMSD below 6.5 Å from native structure. If we take 10, 20, 30, or 40 trajectories and cluster them, we can have on average 44.3, 44.9, 45.6, or 45.7 those cases, respectively.

Received 13 September 2001; Accepted 12 February 2002

approach to a test set of 25 proteins and trace the energy and structures of the proteins at each MC step. We find that, for the same CPU time, parallel hyperbolic sampling (PHS) can identify much lower-energy structures and cover a larger region of structure phase space than replica sampling can. As a comparison with our previous work,[7] we use the new methodology on the same benchmark set of 65 proteins. We successfully predict 50 cases having at most five clusters, at least one whose RMSD from the native state is below 6.5 Å, whereas the CPU time is significantly reduced because here only one trajectory is needed.

By calculating the distance between the structure clusters produced by both PHS and RS simulations, we find that there exists a strong correlation between the quality of predicted structures and the minimum relative RMSD between structure clusters produced by different search engines. This correlation in mrRMSD may be useful in blind prediction as an indicator of the likelihood of successful fold prediction.[9] As a confirmation, we show that in our fold experiment, the prediction power of choosing the common cluster from PHS and RS samples is more favorable than that of choosing the lowest-energy cluster obtained from PHS alone.

MATERIALS AND METHODS

The basic idea of PHS is to apply a nonlinear transformation to the energy E

$$\tilde{E} = \begin{cases} \operatorname{arcsh}(E - E_0), & E \geq E_0, \\ -\infty, & E < E_0, \end{cases} \quad (1)$$

where E_0 is the protein energy of the current structure, and arcsh is the inverse hyperbolic sine function. Here and in the future, the energy is scaled by a unit of ϵ_0, which serves to define the dimensionless temperature T in our calculations.

Obviously, the effective potential $\tilde{E}$ is dynamically changed for each step of the simulation. However, this transformation does preserve the locations of all minima on the real energy surface E. As shown in Figure 1, the inverse hyperbolic sine transformation does not significantly modify the height of the lower energy barriers over which Metropolis sampling can jump with reasonable ease. This feature is important because it allows the simulation to sample nearby energy regions with comparable efficiency as in canonical Metropolis sampling. However, for higher-energy barriers, the barrier height is logarithmically lowered; this can significantly reduce the trapping time in one local basin. Thus, a simulation based on Eq. 1 is actually equivalent to a Metropolis implementation on a much less rough energy landscape.

The idea of transforming energy landscapes is not new.[4,10–13] For example, in the "diffusion equation method," Piela et al.[10] and Wawak et al.[13] produced a smoothed energy surface by solving the diffusion equation, with the original energy function as the initial condition; in the "Liouville equation method," Ma et al.[11] and Andricioaei and Straub[12] coarse grained the potential surface by integrating the energy function over a Gaussian phase packet. In both approaches, the positions of the surviving local energy minima in the deformed surface, which are controlled by a deformation parameter, usually differ from the minima positions in the undeformed surface. Thus, a multiple iterative procedure is required to gradually lower down the deformation parameter and trace back the minima of the original energy surface.[14] In our transformation of Eq. 1, however, the energy surface deformation is based on the instantaneous energy values, and the positions of all the local energy minima are the same as that of the bare energy function. Therefore, the additional reversing process to recover the undeformed minima is not required. We have also tried various other formats of transformations in our calculations. But it appears that the hyperbolic sine function works best among all our attempts.

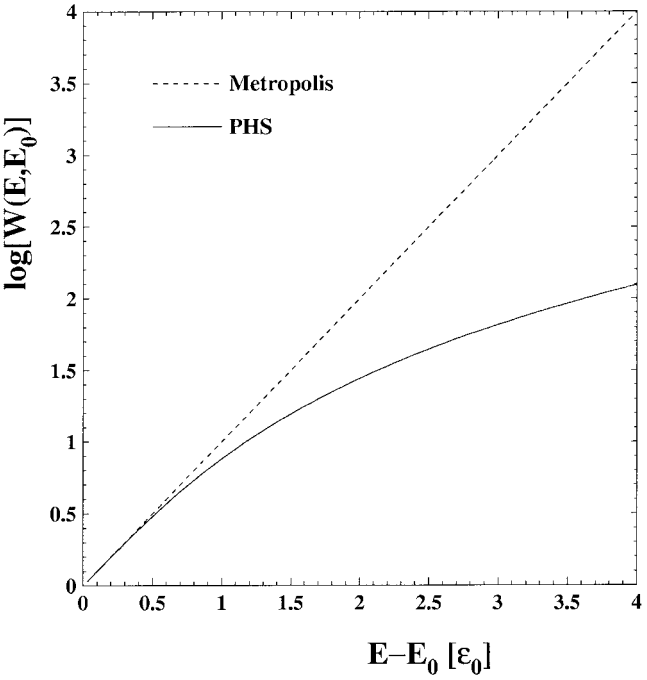

Fig. 1. Comparison of the hyperbolic weight factor and the Boltzmann weight factor as a function of the height of the energy barriers. Here $w(E,E_0)$ is defined as the transition probability from a state of energy E_0 to another state of energy E in the Markov process; see Eq. 2.

The idea of parallel sampling here is similar to those previously used[5,6,15] in which an artificial ensemble consisting of M noninteracting replicas are considered, each at a distinct and fixed temperature. Two sets of movements are then taken into account in our simulations:

(i) Local movements in each replica, which consist of single residue "kink" moves, chain-end moves, two-residue moves, and small "rigid-body" displacements of a larger portion of the model chain.[8] For a protein sequence comprised of N residues, a single time unit consists of N attempts at kink moves, two attempts at chain-end moves, $N-1$ attempts at two-bond moves, and one attempt at a randomly selected large frag-

ment displacement. Each trial movement is accepted or rejected according to:

$$w(E,E_0) = \exp(-\beta\Delta\tilde{E}) = \begin{cases} \exp[-\beta\mathrm{arcsh}(E-E_0)], & E \geq E_0, \\ 1, & E < E_0, \end{cases} \quad (2)$$

where $\beta = 1/k_B T$ is the general inverse temperature. However, before any energy computation, the test for excluded volume violation is always performed, and a trial conformation that would lead to steric collisions of chain units is rejected. Also, conformations with nonphysical distances between two consecutive side-chain units are a priori rejected.

(ii) Swap movements of global conformations between two replicas (say, i and j). The acceptance probability of each swap is:

$$p_{i\leftrightarrow j} = \exp[(\beta_i - \beta_j)(E_i - E_j)]. \quad (3)$$

Although it is not necessary to restrict the swap to the pairs of replicas associated with neighboring inverse temperatures β_i and β_{i+1}, this choice will be optimal because the acceptance ratio will decrease exponentially with the difference $\Delta\beta = \beta_i - \beta_j$.

To optimize the power of the Monte Carlo algorithm, the following important parameters should be carefully selected:

First, the highest simulation temperature T_{high} should be high enough to enable the simulation to jump over all the energy barriers on the artificial landscape defined by Eq. 1 with reasonable ease. The lowest temperature T_{low} should be low enough so that the simulation can scan low-energy basins in sufficient detail. A qualitative goal for the choice of T_{low} is to make the single Markov process slightly trapped. We find that the trap status of a single replica simulation is sensitive to the length of a protein sequence. For longer sequences, the depth of the energy basin is usually deeper than that of shorter proteins. Hence, T_{low} of longer protein sequences should be slightly higher than that for shorter ones.

The number of replicas N_{rep} is another relevant parameter for the implementation. N_{rep} should be large enough so that the replicas at adjacent temperatures are near enough to maintain communication with each other, that is, swaps can occur with reasonable ease.

The third parameter is N_{swap}, the number of local movements between two consecutive global swap movements between different temperatures. The general criterion for choosing the swapping frequency is to follow the higher-temperature simulations until they jump over different energy basins. This timescale can be quantitatively estimated by the integrated autocorrelation time τ of the Markov process.[16]

Actually, all of these three parameters are interrelated and are dependent on considered systems. In our SICHO model system, we have performed a number of initial runs on a set of 13 test proteins (i.e., those marked with * in Table II) and optimized all three parameters, the results of which are summarized in Table I. We show in Figure 2 the histograms of energies of all N_{rep} replicas for 1mba_, the longest protein sequence in our protein set. Indeed, the low-temperature replicas mainly explore the low-energy structures of the protein system, and the high-temperature processes serve to transform the composite ensemble from one region of the energy phase space to another. The intermediate temperatures replicas serve to facilitate communication between these high- and low-temperature replicas.

TABLE I. Monte Carlo Parameters of Parallel Hyperbolic Sampling†

	$T_{low}[\varepsilon_0/k_B]$	$T_{high}[\varepsilon_0/k_B]$	N_{swap}	N_{rep}
$N < 50$	0.2	1.3	300	40
$50 \leq N \leq 100$	0.3	1.3	300	40
$N < 100$	0.4	1.3	500	40

†The parameters are optimized according to a number of test runs on the 13 proteins listed in Table II. N is the number of residues of the calculated sequences.

RESULTS AND DISCUSSION

Throughout this article, we apply the proposed algorithm to the SICHO protein model.[8] In this model, the conformation of the protein chain is specified by the coordinates of the center of mass of the side-chains and the backbone α-carbons. These interaction centers are located on an underlying three-dimensional cubic lattice system with a lattice spacing of 1.45 Å. Depending on the identity of two consecutive residues, the associated main-chain conformation and the rotameric state of the side-chain, the virtual bonds are of variable lengths ranging from 4.35 to 7.94 Å . This covers the distribution seen in real proteins with good fidelity. There are 646 allowed bond vectors. With this geometric presentation, all PDB structures could be represented with an average RMSD of about 0.8 Å.[18]

The force field of our SICHO protein model consists of three types of terms. The first terms are sequence-independent contributions that provide biases to regular secondary structures, penalties on nonprotein-like conformations, hydrogen-bond interactions, and a centrosymmetric potential. The second terms are sequence-specific contributions that consist of a weak bias toward the predicted secondary structure, a sequence-dependent short-range geometric bias for fragments, and a protein-specific pairwise potential. Finally, there are the tertiary restraints for long-range contacts and short-range distances, which are derived from threading and multiple-sequence alignments.[18] In all these analyses, those PDB[19] structures whose sequences are similar to the objective 65 test sequences have been removed from the structural database (at greater than 25% sequence identity). Detailed descriptions and analysis of the construction of the model force field have been recently published.[7,8,18,20,21]

Minimum Energy and RMSD of 25 Test Proteins

We first apply the PHS algorithm to a test set of 25 proteins that cover a range of lengths from 44 to 146

TABLE II. Comparison of Low-Energy States and Minimum RMSD From PHS and RS[†]

		$\langle E(T_1)\rangle$		$E_{\min}$		$\langle E_{\min}\rangle_{20}$		$\mathrm{RMSD}_{\min}$	
ID	Length	RS	PHS	RS	PHS	RS	PHS	RS	PHS
1fc2C	44	−354.7	−364.5	−397.2	−412.1	−386.5	−401.2	2.802	2.628
*1gpt_	47	−403.2	−418.6	−454.5	−454.6	−441.3	−441.4	2.139	1.993
*1tfi_	50	−396.2	−427.9	−445.5	−464.2	−436.0	−455.6	2.865	2.834
1bq9A	53	−427.9	−458.0	−482.1	−487.0	−466.7	−472.5	3.927	3.899
1vif_	60	−447.0	−471.3	−510.1	−518.5	−489.1	−503.1	3.013	2.554
*1fas_	61	−605.1	−645.7	−664.3	−688.7	−650.3	−676.5	3.341	2.589
1ctf_	68	−571.9	−594.0	−636.9	−647.3	−607.0	−632.7	8.643	8.820
*1ftz_	70	−608.8	−621.7	−687.4	−689.7	−661.4	−667.2	4.307	4.103
*1ah9_	71	−645.8	−671.0	−711.4	−726.9	−690.2	−710.0	7.181	4.435
1lea_	72	−739.4	−765.4	−809.2	−818.5	−788.7	−799.5	2.141	2.108
1kjs_	74	−588.4	−603.8	−663.5	−666.0	−632.2	−656.7	4.790	4.553
*1ner_	74	−622.9	−624.6	−698.8	−699.9	−649.4	−659.8	4.368	4.354
*1a32_	78	−700.5	−701.8	−803.1	−776.8	−753.4	−747.6	10.381	10.546
*1aoy_	78	−796.4	−827.7	−868.0	−873.7	−846.6	−849.0	3.457	3.127
1wiu_	93	−1036.8	−1126.3	−1117.6	−1184.5	−1090.2	−1148.3	2.845	2.638
2ezk_	93	−718.8	−735.1	−779.5	−829.4	−750.3	−803.0	5.344	1.844
1tsg_	98	−724.3	−754.2	−797.2	−817.9	−781.5	−798.9	8.795	8.525
1ksr_	100	−896.8	−967.8	−987.7	−1032.8	−954.1	−1005.3	4.241	3.087
*2lfb_	100	−773.3	−778.0	−864.1	−872.4	−841.9	−844.0	7.548	7.557
1tlk_	103	−1207.6	−1250.5	−1285.9	−1320.5	−1261.4	−1282.0	2.525	1.256
*1hmdA	113	−1144.8	−1219.8	−1262.6	−1311.3	−1211.5	−1283.6	2.700	2.632
*1pdo_	121	−1131.4	−1141.7	−1238.6	−1244.2	−1209.0	−1229.6	4.896	4.244
*4fgf_	121	−1008.5	−1048.3	−1112.2	−1144.6	−1079.7	−1111.9	7.862	6.726
1hlb_	138	−1464.2	−1642.9	−1682.2	−1758.8	−1544.7	−1702.7	4.330	2.891
*1mba_	146	−1430.9	−1658.1	−1655.7	−1813.3	−1512.1	−1713.5	4.475	2.726
Average	85	−777.8	−820.7	−864.6	−890.1	−829.4	−863.8	4.757	4.107

[†]In each column, the values on the left are obtained by replica sampling (RS) and those on the right by the parallel hyperbolic sampling (PHS) method. Each MC run consists of 40 replicas and 200 MC sweeps in each replica. The unit of energy is in ε_0 and the RMSD is in angstroms. The proteins marked with * have been used to tune the Monte Carlo parameters found in Table I.

residues (see Table II). Two hundred MC sweeps are performed for each protein, each sweep consisting of N_{swap} units of local movements in each replica. In each update of local movement, we trace and record the protein's energy of every configuration and the RMSD from its X-ray structure. As a comparison, we also perform normal replica sampling simulations on the proteins for the similar CPU time, where all the MC parameters of the RS simulations have been optimized in previous articles.[8,21] In columns 3 and 4 of Table II, we calculate the average energy of the lowest-temperature replica [$E(T_1)$], because low-energy structures are usually explored by this replica, according to Figure 2. In columns 5 and 6, we trace and record the minimum energy $E_{\min}$ that the simulations of all replicas ever reach. In columns 7 and 8, we divide the whole sample into 20 subsamples and record the minimum energy separately in each subsample. $[E_{\min}]_{20}$ denotes the average of the minimum energies found in these 20 subsamples. Except for 1a32_, which does not fold to the native state in our model, PHS can find lower-energy structures than RS in all the cases.

In columns 9 and 10 of Table II, we list also the lowest RMSD from the native structure ever found in both simulations. Again, smaller RMSD structures can be found by parallel hyperbolic sampling in most cases. This indicates that PHS can explore a larger phase space in the same amount of computation time, compared with the regular replica sampling.

As an example, we compare in Figure 3 the time series of the energy and RMSD of 1mba_ sequence of the lowest-temperature replica, produced, respectively, by PHS and by the RS simulations. Obviously, the RS simulation is stuck in some higher-energy regions and its structure fluctuates in a region of phase space that is quite far away from the native state.

Structure Prediction of 65 Benchmark Proteins

In our prediction experiment, we take the same set of 65 proteins as used previously.[7] Among this protein set, there are 4 small proteins with little secondary structure, 21 α-helical proteins, 20 β-sheet proteins, and 20 α/β proteins, following the CATH classification.[22] The proteins range in length from 39 to 146 amino acids.

For each protein, we perform a number of Monte Carlo runs based on the PHS algorithm, each run starting from different initial random numbers. We set up a structure pool by picking structures from snapshots of MC processes. In principle, the time interval between selected neighboring snapshots should be long enough so that they are structurally uncorrelated. This approximately corresponds to the integrated autocorrelation time τ ($\sim N_{\mathrm{swap}}$).[16] In our case, we pick up one snapshot after each MC sweep

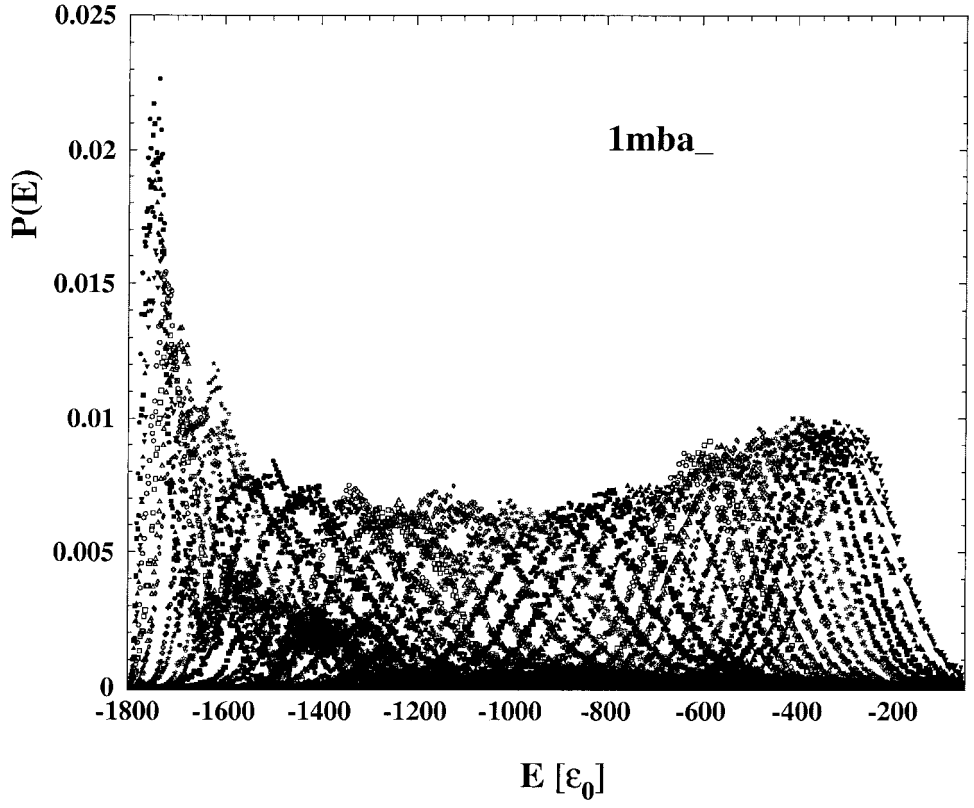

Fig. 2. Energy histogram of different replicas of a representative run of parallel hyperbolic sampling on 1mba_. Two hundred MC sweeps are performed in each replica.

(i.e., N_{swap} units of local movements. Because the near-native structures are most likely explored in low-temperature replicas, we only pick structures for the lowest-temperature replica. In each Monte Carlo run, we perform 800 sweeps. Therefore, there are 800 structures in each trajectory. A structure pool can include more than one trajectory depending on the number of performed Monte Carlo runs.

Because of the imperfections in the current potential, usually we cannot reliably obtain the native fold by choosing the lowest-energy structure. Having in mind that, for a reasonable force field, the partition function of a near-native state should be significantly larger than that of non-native states; thus, here we look for the common folds by clustering all the configurations in our structure pool and choosing the centroid of the selected cluster. To determine the degree of convergence, clustering is done in two steps. First, the clustering is done in each trajectory, and then the obtained centroids in different trajectories are again clustered.[23] If there is only one trajectory in the structure pool, only the first step of clustering is needed.

As a test, we first perform 50 PHS Monte Carlo runs on each of a small set of test proteins (see Fig. 4). We then subject different number of trajectories to the clustering process. Figure 4 shows the clustering result as a function of the number of subjected trajectories. Although the total number of resulting clusters increases with the number of trajectories, the quality of the clusters (i.e., the smallest RMSD of the produced clusters) does not improve with an increasing number of Monte Carlo runs. Our unpublished data show the RS simulation does not have this feature, and at least 50 trajectories are needed to obtain the best RMSD values.[†] This may indicate that the thermalization of PHS simulation is much faster so that only one PHS trajectory is sufficient to explore all the important areas of conformational space. In other words, the best structure can be safely obtained in each PHS simulation within a sufficient simulation time. On the basis of this finding, we will run only one trajectory for each protein in the following calculations.

In column 5 of Table III, we show the predicted results of the PHS simulation. The shown value is the smallest RMSD from the native structure among all the cluster centroids. The first value in parentheses is the order number of the best cluster, where the clusters have been ordered according to the average energy of the structures within the cluster and the second value in parentheses is the total number of produced clusters. In column 4, we also show the predicted results of the RS simulation of 50 trajectories, which is taken directly from our previous article.[7] In the RS simulation, there are 46 cases among all 65 proteins, in which the RMSD of the best cluster (in the

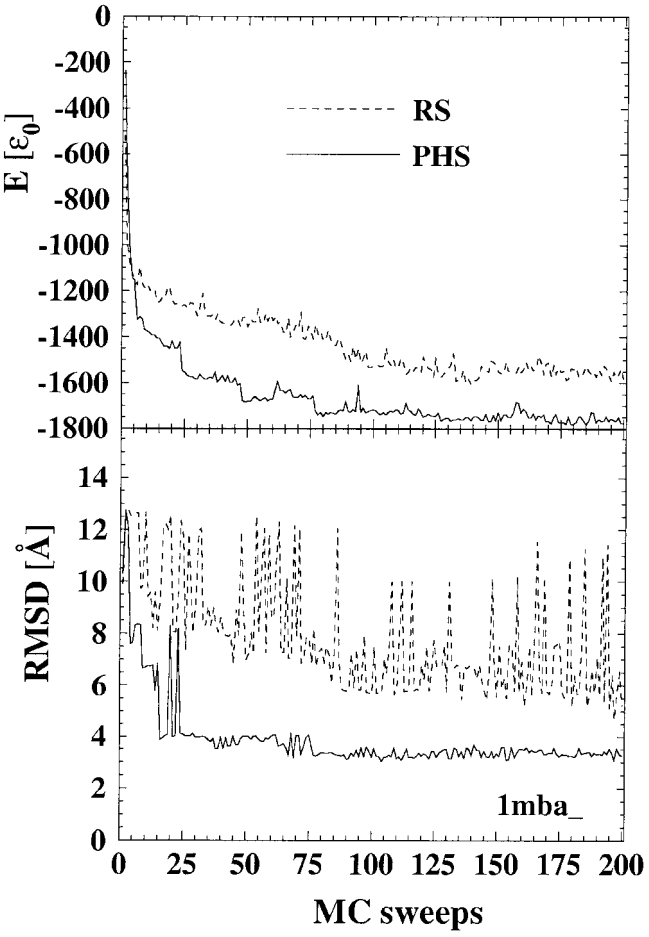

Fig. 3. Comparison of the time series of the energy and RMSD from native of the lowest-temperature replica of 1mba_ obtained by replica sampling (dashed lines) and by parallel hyperbolic sampling (solid lines). Two hundred Monte Carlo sweeps are performed in each case.

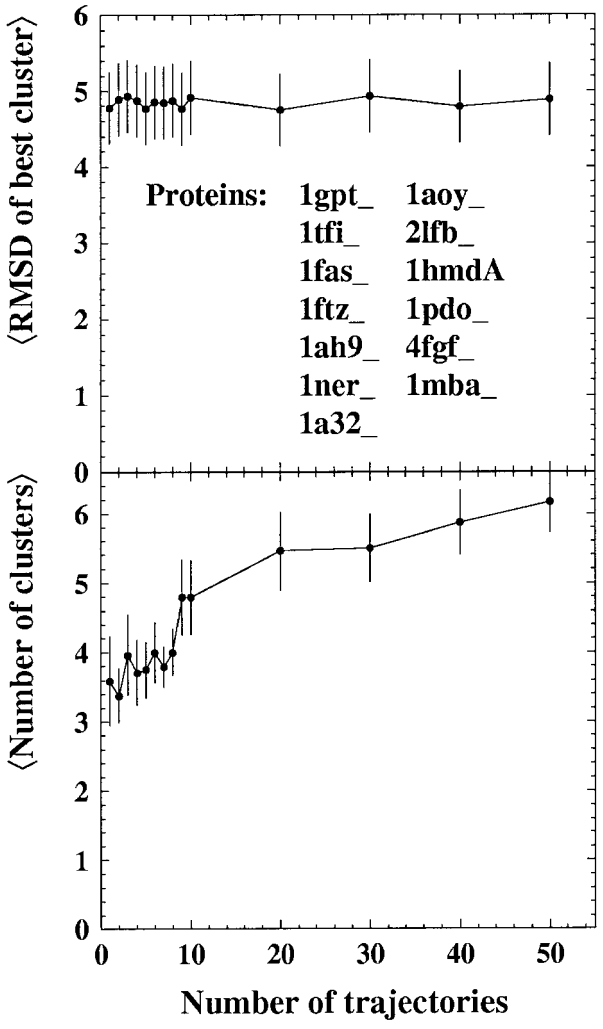

Fig. 4. RMSD of the best cluster from native and the total number of clusters as a function of the number of trajectories run using the parallel hyperbolic sampling algorithm. The data are taken from the average of the shown proteins. Each trajectory includes 800 Monte Carlo sweeps.

top five clusters) is below 6.5 Å. This number increases to 50 when PHS is used. By experience, when the RMSD of a predicted structure is less than 6.5 Å, the topology of the protein structure is generally correct. This means that, by using the same potential, PHS can successfully predict more protein structures (4 of 65), whereas the CPU cost is less. We have also set the threshold of "successful" RMSD from native to 6.0, 5.0, 4.0, and 3.0 Å, respectively. PHS can obtain a higher number of successful predictions in all the cases (see the last five rows of Table III).

In Figure 5, we show four examples in which the PHS approach has significantly improved the quality of the calculated structure of the RS simulation. The RMSD from native of these predicted structures obtained by the RS method are, respectively, 10.2 Å (2ezk_), 8.0 Å (1stu_), 5.4 Å (1tlk_), and 8.6 Å (2af8_); the corresponding values calculated by the PHS approach are, respectively, 4.5 Å (2ezk_), 5.7 Å (1stu_), 3.1 Å (1tlk_), and 4.4 Å (2af8_). None of these four cases belongs to the test set of 13 proteins on which we took initial runs for tuning the MC parameters of the PHS simulations.

Common Structure by Different Search Engines

In limited CPU time, different search engines may explore different parts of structure phase space. It is of interest to check whether these two simulations using PHS and RS search common structure clusters and whether the identification of the common structures can give any hint to the quality of a blind fold prediction. One way of attacking this problem is to calculate the relative root-mean-square deviation (rRMSD) between these two sets of clusters and identify the common fold by looking for the cluster pair with the minimum rRMSD (mrRMSD).

In Figure 6, we calculate the correlation between the calculated RMSD from native and the mrRMSD by PHS and RS simulations. To reduce the statistical error, we average the calculated RMSD over those proteins whose mrRMSD stays in the range ($mrRMSD_0-0.8$, $mrRMSD_0=0.8$]. The data in Figure 6 show an obvious

TABLE III. Summary of Prediction Results on 65 Benchmark Proteins†

						Comparison of RS and PHS	
ID	Structural type	Length	RS	PHS	Lowest energy	50 RS samples	1 RS sample
1a32_	α	85	7.4 (1/4)	6.4 (3/3)	8.0	8.0/2.4 (1)	8.0/2.3 (1)
1ah9_	β	71	7.5 (7/8)	6.8 (1/3)	6.8*	6.8/4.9 (1)*	10.8/6.9 (2)
1aoy_	α	78	4.5 (1/5)	4.4 (1/3)	4.4*	4.4/1.4 (1)*	4.4/1.4 (1)*
lbq9A	β	53	6.9 (1/8)	6.4 (7/8)	9.7	9.7/4.3 (1)	9.7/3.9 (1)
1bw6A	α	56	5.0 (1/7)	7.2 (1/3)	7.2*	11.5/2.4 (2)	7.2/2.6 (1)*
1c5a_	α	66	5.8 (3/6)	4.8 (1/3)	4.8*	4.8/1.6 (1)*	4.8/3.0 (1)*
1cewI	α/β	108	7.2 (1/5)	5.8 (1/3)	5.8*	5.8/4.2 (1)*	5.8/5.0 (1)*
1cis_	α/β	66	4.8 (2/6)	4.4 (2/3)	4.4	4.4/3.9 (2)*	4.4/3.8 (2)*
1csp_	β	64	3.6 (1/7)	4.7 (1/3)	4.7*	4.7/3.7 (1)*	4.7/3.7 (1)*
1ctf_	α/β	68	9.6 (2/5)	10.2 (2/2)	12.0	12.0/2.6 (1)	12.0/2.2 (1)
lerv_	α/β	105	2.3 (1/2)	2.2 (1/4)	2.2*	2.2/1.1 (1)*	2.2/1.2 (1)*
1fas_	β	61	3.4 (1/3)	3.7 (1/2)	3.7*	3.7/1.6 (1)*	3.7/1.6 (1)*
1fc2C	Small	44	3.6 (2/2)	3.4 (2/2)	7.0	7.0/1.0 (1)	7.0/1.1 (1)
1ftz_	α	70	2.9 (1/2)	2.9 (1/3)	2.9*	2.9/0.6 (1)*	2.9/0.6 (1)*
1gpt_	α/β	47	3.4 (2/4)	2.6 (1/1)	2.6*	2.6/1.7 (1)*	2.6/1.7 (1)*
1hlb_	α	138	2.6 (1/9)	2.7 (1/4)	2.7*	2.7/0.4 (1)*	2.7/0.5 (1)*
1hmdA	α	113	2.6 (1/5)	2.4 (1/1)	2.4*	2.4/0.9 (1)*	2.4/1.0 (1)*
1hp8_	α	68	4.9 (1/2)	4.5 (1/3)	4.5*	4.5/0.9 (1)*	4.5/1.0 (1)*
1ife_	α/β	91	6.3 (3/10)	8.7 (1/2)	8.7*	8.7/3.9 (1)*	8.7/3.7 (1)*
1ixa_	Small	39	4.5 (2/7)	4.1 (2/5)	7.6	8.2/2.3 (3)	8.2/2.4 (3)
1iyv_	β	79	10.6 (3/11)	8.5 (2/2)	9.1	9.1/8.9 (1)	8.5/8.3 (2)*
1kjs_	α	74	4.5 (1/6)	4.8 (1/2)	4.8*	4.8/1.5 (1)*	4.8/1.7 (1)*
1ksr_	β	100	5.1 (1/9)	4.7 (1/1)	4.7*	4.7/2.3 (1)*	4.7/2.1 (1)*
1lea_	α	72	3.7 (1/5)	3.2 (1/1)	3.2*	3.2/1.4 (1)*	3.2/1.4 (1)*
1mba_	α	146	2.7 (1/3)	2.4 (1/4)	2.4*	2.4/1.3 (1)*	2.4/1.9 (1)*
1ner_	α	74	4.1 (1/6)	4.5 (4/5)	5.8	10.3/1.9 (3)	10.8/3.5 (2)
1ngr_	α	85	2.7 (1/3)	2.4 (1/3)	2.4*	2.4/1.7 (1)*	2.4/2.1 (1)*
1nkl_	α	78	3.0 (1/5)	3.4 (1/3)	3.4*	3.4/1.7 (1)*	3.4/1.8 (1)*
1nxb_	β	53	3.6 (3/3)	4.9 (2/2)	6.8	4.9/1.2 (2)*	4.9/2.7 (2)*
1pdo_	α/β	121	6.5 (2/2)	6.6 (2/7)	9.1	9.1/2.4 (1)	9.1/2.0 (1)
1pgx_	α/β	56	2.3 (1/4)	2.2 (1/8)	2.2*	2.2/0.9 (1)*	2.2/0.8 (1)*
1poh_	α/β	85	3.3 (1/5)	2.7 (2/2)	10.4	2.7/1.0 (2)*	2.7/1.2 (2)*
1pou_	α	71	3.7 (1/10)	3.3 (1/4)	3.3*	3.3/2.1 (1)*	3.3/2.1 (1)*
1pse_	β	69	8.4 (4/6)	8.1 (2/9)	12.9	8.1/2.6 (2)*	10.1/3.3 (3)
1rip_	β	81	9.3 (5/21)	11.1 (1/6)	11.1*	11.2/2.9 (3)	11.2/2.5 (3)
1rpo_	Small	61	3.7 (4/4)	9.6 (3/3)	10.9	10.9/0.7 (1)	10.9/0.1 (1)
1shaA	α/β	103	3.6 (1/14)	3.3 (1/1)	3.3*	3.3/2.4 (1)*	3.3/2.4 (1)*
1shg_	β	57	4.9 (1/8)	4.3 (1/1)	4.3*	4.3/4.1 (1)*	4.3/4.3 (1)*
1sro_	β	66	6.4 (2/6)	5.4 (2/3)	9.3	5.4/4.5 (2)*	5.4/3.9 (2)*
1stfI_	α/β	98	7.1 (5/5)	5.7 (3/6)	12.6	5.7/3.1 (3)*	5.7/2.6 (3)*
1stu_	α/β	68	8.0 (4/10)	5.7 (3/3)	10.0	8.2/2.9 (2)	8.2/2.5 (2)
1tfi_	β	50	4.3 (4/5)	4.5 (2/2)	8.3	8.3/1.1 (1)	4.5/1.6 (2)*
1thx_	β	108	2.2 (1/5)	2.7 (1/1)	2.7*	2.7/1.6 (1)*	2.7/1.6 (1)*
1tit_	β	89	2.4 (1/3)	2.0 (1/5)	2.0*	2.0/1.4 (1)*	2.0/1.3 (1)*
1tlk_	β	103	5.4 (1/2)	3.1 (2/2)	4.1	3.1/4.7 (2)*	3.1/5.1 (2)*
1tsg_	α/β	98	8.7 (1/7)	9.9 (3/17)	11.6	13.1/8.7 (12)	9.9/8.3 (3)*
1ubi_	α/β	76	3.6 (1/8)	2.9 (2/2)	3.8	3.8/0.4 (1)	3.8/0.5 (1)
1vcc_	α/β	77	9.9 (1/6)	8.4 (1/9)	8.4*	10.8/5.3 (4)	12.0/5.7 (9)
1vif_	β	60	4.4 (2/12)	3.4 (1/3)	3.4*	3.4/3.2 (1)*	3.4/3.1 (1)*
1wiu_	β	93	2.6 (1/3)	2.7 (1/2)	2.7*	2.7/1.4 (1)*	2.7/1.3 (1)*
256bA	α	106	3.4 (1/3)	3.7 (1/1)	3.7*	3.7/1.8 (1)*	3.7/1.6 (1)*
2af8_	α	86	8.6 (8/10)	4.4 (1/6)	4.4*	11.5/7.5 (4)	11.5/8.8 (4)
2azaA	β	129	4.5 (1/3)	3.7 (1/2)	3.7*	3.7/3.7 (1)*	3.7/4.2 (1)*
2bby_	α	69	4.9 (1/5)	4.1 (1/3)	4.1*	4.1/3.2 (1)*	4.1/3.4 (1)*
2ezh_	α	65	5.2 (2/6)	4.6 (2/3)	6.8	4.6/3.2 (2)*	6.8/6.6 (1)
2ezk_	α	93	10.2 (7/8)	4.5 (6/6)	12	13.8/4.6 (4)	13.8/4.7 (4)
2fdn_	α/β	55	9.6 (4/10)	8.7 (7/7)	10.8	10.4/5.7 (2)	10.4/5.6 (2)
2fmr_	α/β	65	3.7 (1/2)	3.8 (1/2)	3.8*	3.8/1.0 (1)*	3.8/1.2 (1)*
2lfb_	α	100	4.9 (9/10)	5.6 (8/9)	11.3	10.1/1.6 (7)	9.8/3.0 (2)
2pcy_	β	99	4.0 (1/4)	3.0 (1/2)	3.0*	3.0/3.1 (1)*	3.0/2.8 (1)*

TABLE III. (Continued)

ID	Structural type	Length	RS	PHS	Lowest energy	Comparison of RS and PHS	
						50 RS samples	1 RS sample
2ptl_	α/β	60	2.5 (1/3)	2.9 (1/3)	2.9*	2.9/1.0 (1)*	8.8/1.1 (2)
2sarA	α/β	96	4.1 (1/6)	4.3 (1/2)	4.3*	4.3/2.5 (1)*	12.1/3.4 (2)
4fgf_	β	121	9.7 (1/5)	9.3 (2/3)	12.1	12.1/4.8 (1)	12.1/5.9 (1)
5fd1_	α/β	106	9.7 (4/5)	10.3 (3/6)	13.3	10.3/8.6 (3)*	13.7/8.5 (2)
6pti_	Small	57	7.2 (5/7)	6.6 (1/2)	6.6*	9.2/2.5 (2)	9.2/3.9 (2)
Average		80.6	5.25 (2.1/6.0)	4.99 (1.8/3.6)	6.21	6.06/2.77	6.29/3.02
Total numbers:							
RMSD < 6.5			46	50	37	40	38
RMSD < 6.0			44	48	37	40	38
RMSD < 5.0			40	43	35	37	35
RMSD < 4.0			25	26	23	25	24
RMSD < 3.0			11	15	13	14	13

†In columns 4 and 5 are the RMSD of the best cluster centroid from the native structure calculated by RS and PHS simulations, respectively. The first number in parentheses denotes the order number of the best cluster, and the second number in parentheses is the total number of produced clusters. In column 6 is the predicted RMSD if we choose the first cluster (i.e., the lowest-energy cluster in PHS simulation). In column 7 is the predicted RMSD if we choose the PHS cluster that has the mrRMSD from the clusters formed in the 50 RS trajectories; the second value is the mrRMSD, and the number in parentheses denotes the order number of the chosen clusters. Column 8 is similar to column 7, but RS clusters are produced in one randomly chosen trajectory. In columns 4–8 denotes the best cluster is chosen in given approach. The last five rows show the number of proteins whose predicted RMSD is below the respective threshold values shown in column 1. The units of both RMSD and mrRMSD are in angstroms.

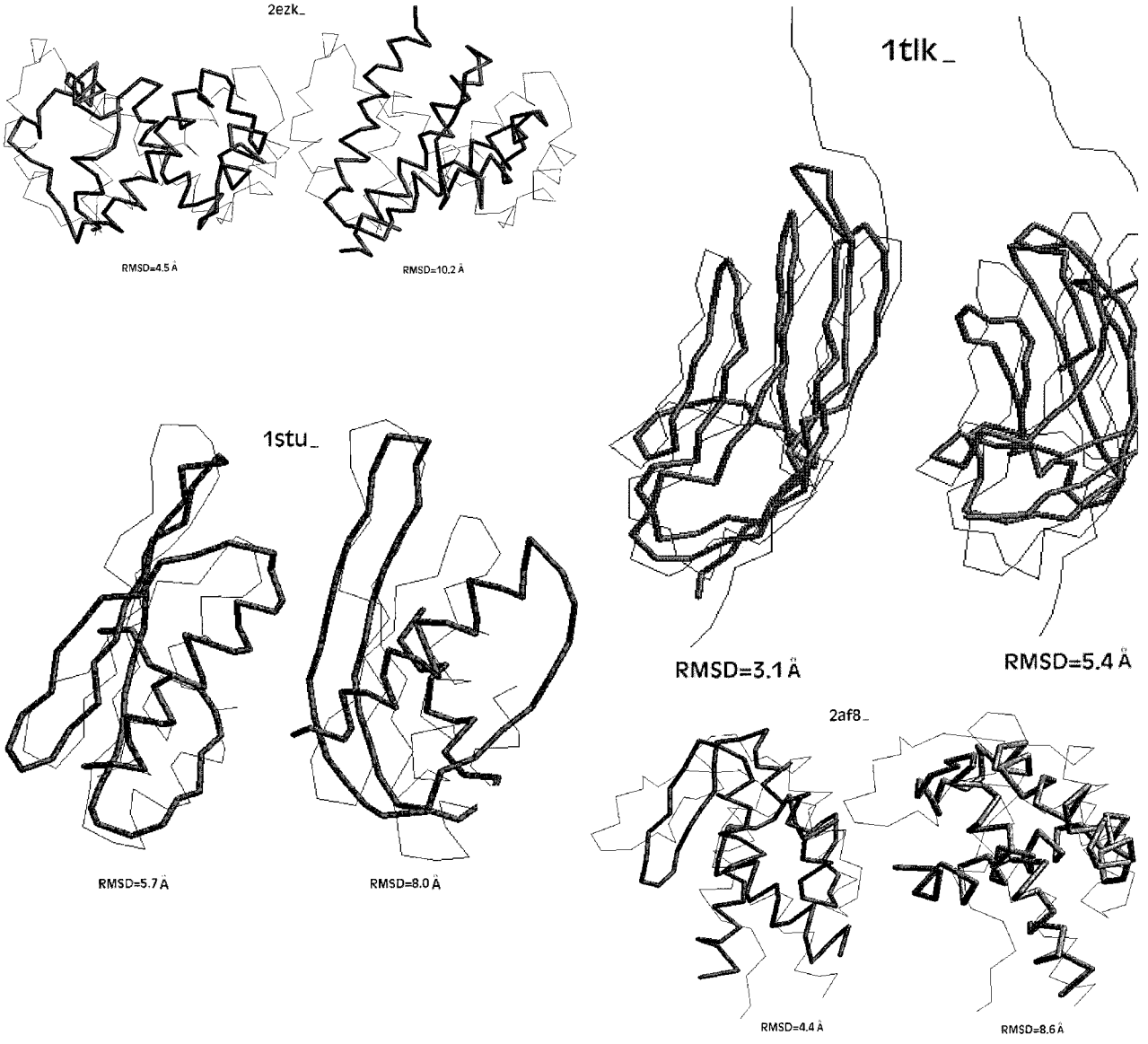

Fig. 5. Structures of four representative examples in which the quality of the predicted structures is significantly improved by parallel hyperbolic sampling (**left**) compared with that obtained by replica sampling (**right**). The backbone of predicted structures are shown in thick lines, and that of native structures in thin lines.

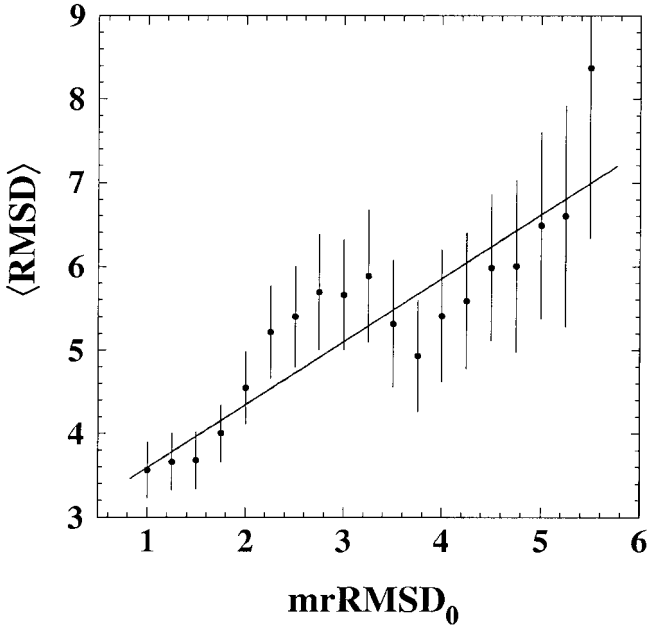

Fig. 6. The correlation between the RMSD of the predicted structure from native obtained by PHS and the mrRMSD of the clusters produced by PHS from the clusters produced by RS simulations. The average RMSD is taken on those proteins whose mrRMSD is in the (mrRMSD$_0$−0.8, mrRMSD$_0$+0.8). For the last point, however, the average is done over the proteins whose mrRMSD in (mrRMSD$_0$, +∞]. The line is the minimum mean square fit of Eq. 4 with a correlation coefficient $c = 0.76 \pm 0.11$.

correlation of the average RMSD and mrRMSD$_0$. By fitting the data to the equation

$$\langle \mathrm{RMSD} \rangle = c^* \mathrm{mrRMSD}_0 + b \quad (4)$$

We obtain $c = 0.76 \pm 0.11$. This correlation indicates that, when a common fold can be found by different search engines, the predicted fold is more likely to be successful than the cases when no common fold is found. In this context, the value of mrRMSD by different search engines may be considered to be a possible indicator of the likelihood of successful fold predictions.

Because the cluster with the best structure(s) does not always correspond to the cluster of the lowest energy, the correct identification of the closest to native is a nontrivial issue. In the PHS simulation, there are on average 3.6 clusters for each protein, and the average order number of the best cluster is 1.8. These two numbers are 6.0 and 2.1 in the RS simulations, respectively. Obviously, this advantage of PHS sampling is important because it increases the probability of the identification of the best structure from the clusters.

In our simulations, we can identify the best cluster closest to the native in 39 cases from 65 benchmark proteins if we choose the lowest-energy cluster. As a proof of the above-mentioned correlation, we find that choosing the cluster of mrRMSD in the PHS sample is more favorable than choosing the lowest-energy cluster. For example, if we compare the PHS and RS samples of all the 50 trajectories and choose the cluster of mrRMSD, we can identify the best cluster in 43 cases (column 7 in Table III). If we take randomly one trajectory from replica sampling and do the comparison with PHS, we can identify the best cluster in 41 cases (column 8 in Table III). As shown in the last five rows of Table III, in both cases, the numbers of successful predictions by the comparison are also slightly higher than that of choosing the lowest-energy cluster.

CONCLUSIONS

In this work, we have extended replica sampling Monte Carlo method through the local flattening of the high-energy barriers by an inverse hyperbolic sine function, thus allowing a single replica simulation to more quickly explore the low-energy basins of the protein's energy landscape. We have applied the proposed algorithm to the simulation of the SICHO protein model and find that PHS significantly reduces the CPU time required to effectively sample conformational space compared with regular replica sampling. After clustering the produced structures, we can successfully predict, among 65 test proteins, 50 cases in which at least one of the top five clusters has a RMSD from the native structure below 6.5 Å compared with 46 cases using canonical replica sampling requiring more CPU time. This is also at least partially suggestive that the force field is not as poor as we originally thought; with better sampling, better structures are obtained on average, an encouraging result.

To identify the correct near-native structure in the simulation, we calculate the mrRMSD between the clustered structures produced by the two distinct sampling schemes PHS and RS. We find that choosing the common cluster by different search engines can increase the prediction accuracy, compared with that of choosing the lowest-energy cluster. A correlation coefficient of about 0.76 is found between the RMSD of a predicted structure and the mrRMSD. This correlation may be useful in blind fold prediction to use the value of mrRMSD as a possible indicator of the quality of the prediction structures.

It should be mentioned that a drawback of PHS, in contrast with other algorithms such as RS, is that because the energy landscape is dynamically changed in each update, no thermodynamic expectation can be calculated from the simulation. Thus, for systems having a relatively smooth energy landscape or in the presence of a very efficient move set, PHS may not be more efficient than other competing methods. However, in many realistic optimization problems where the thermodynamic behavior is not the main focus of the calculation, parallel hyperbolic sampling can be a useful search protocol to identify the very low energy states. Such is the case here, and we plan to exploit this speed up by predicting the structure of all the small proteins in a number of genomes.

REFERENCES

1. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223–230.
2. Newman MEJ, Barkema GT. Monte Carlo methods in statistical physics. Oxford: Clarendon Press; 1999.

3. Hansmann UHE, Okamoto Y. The generalized-ensemble approach for protein folding simulations. Ann Rev Comp Phys World Scientific, Singapore, 1998;VI:129–157.
4. Wales DJ, Scheraga HA. Global optimization of clusters, crystals, and biomolecules. Science 1999;285:1368–1372.
5. Swendsen RH, Wang JS. Replica Monte Carlo simulation of spin glasses. Phys Rev Lett 1986;57:2607–2609.
6. Hansmann UHE. Parallel tempering algorithm for conformational studies of biological molecules. Chem Phys Lett 1997;281: 140–150.
7. Kihara D, Lu H, Kolinski A, Skolnick J. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. Proc Natl Acad Sci USA 2001;98:10125–10130.
8. Kolinski A, Skolnick J. Assembly of protein structure from sparse experimental data: an efficient Monte Carlo model. Proteins 1998;32:475–494.
9. Kihara D, Zhang Y, Kolinski A, Skolnick J. Ab initio protrin structure prediction on a genomic scale: application to the *Mycoplasma Genitalium*. Proc Natl Acad Sci USA 2002;99:5993–5998.
10. Piela L, Kostrowicki J, Scheraga HA. The multiple-minima problem in the conformational-analysis of molecules-deformation of the potential-energy hypersurface by the diffusion equation method. J Phys Chem 1989;93:3339–3346.
11. Ma J, Hsu D, Straub JE. Approximate solution of the classical Liouville equation using Gaussian phase packet dynamics: application to enhanced equillibrium averaging and global optimization. J Chem Phys 1993;99:4024–4035.
12. Andricioaei I, Straub JE. Global optimization using bad derivatives: derivative-free method for molecular energy minimization. J Comp Chem 1998;19:1445–1455.
13. Wawak RJ, Gibson KD, Liwo A, Scheraga HA. Theoretical prediction of a crystal structure. Proc Natl Acad Sci USA 1996;93:1743–1746.
14. Pillardy J, Liwo A, Groth M, Scheraga HA. An efficient deformation-based global optimization method for off-lattice polymer chains: self-consistent basin-to-deformed-basin mapping (SCB-DBM). Application to united-residue polypeptide chains. J Phys Chem B 1999;103:7353–7366.
15. Gront D, Kolinski A, Skolnick J. Comparison of three Monte Carlo conformational search strategies for a proteinlike homopolymer model: folding thermodynamics and identification of low-energy structures. J Chem Phys 2000;113:5065–5071.
16. Zhang Y, Skolnick J. Parallel-hat tempering: a Monte Carlo search scheme for the identification of low energy structures. J Chem Phys 2001;117:362–367.
17. Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. A method for the improvement of threading-based protein models. Proteins 1999;37: 592–610.
18. Skolnick J, Kihara D. Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. Proteins 2001;42:319–331.
19. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
20. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
21. Kolinski A, Betancourt MR, Kihara D, Rotkiewicz P, Skolnick J. Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. Proteins 2001;44:133–149.
22. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
23. Betancourt MR, Skolnick J. Finding the needle in a haystack: educing native folds from ambiguous ab initial protein structure predictions. J Comp Chem 2001;22:339–353.