Optimizing the Performance of Bias-Exchange Metadynamics: Folding a 48-Residue LysM Domain Using a Coarse-Grained Model

Pilar Cossio,* Fabrizio Marinelli, Alessandro Laio, and Fabio Pietrucci

SISSA- Scuola Internazionale Superiore di Studi Avanzati, and CNR-INFM DEMOCRITOS National Simulation Center, via Beirut 2−4, I-34014 Trieste, Italy

Received: August 3, 2009; Revised Manuscript Received: January 14, 2010

Computer simulation of complex conformational transitions in biomolecules, such as protein folding, is considered one of the main goals of computational chemistry. A recently developed methodology, bias-exchange metadynamics, was successfully used to reversibly fold some small globular proteins. The objective of this work is to further improve this promising technique. This is accomplished by searching for the optimal set of parameters that enable folding a 48 amino acid protein, 1E0G, in the shortest possible time, using a coarse-grained force field UNRES. It is shown that bias-exchange metadynamics, if appropriately optimized, allows finding the folded state of 1E0G significantly faster than normal replica exchange.

I. Introduction

The physical process in which a polypeptide chain folds from a random coil into its characteristic and functional tertiary structure is called protein folding. Decades of experimental work have led to a deep understanding of this process, and a large number of structures have been resolved using X-rays or NMR. Instead, computer-based protein structure prediction is still considered extremely challenging:[1] nowadays, direct simulation can access at most the microsecond time scale, which remains far from those that would be necessary to provide a truly quantitative and statistically reliable description.[2]

Protein structure prediction can be categorized into template-based modeling and free modeling.[3] Template-based methods search for the most likely structure by suitable combinatorial or alignment algorithms exploiting experimentally known folds. These approaches are very successful when the primary sequence has a high similarity with one (or more) proteins of known structure, but they encounter problems when sequence analogs do not exist. In these conditions, the only approach that can be used is free modeling (or "ab initio" folding), where the prediction is done by identifying the most stable state through the evolution of the particles under the action of a suitable cost-function or empirical potential. In some popular and successful approaches, such as ROSETTA[4] and TASSER,[5] the cost-function is constructed in a template-like manner using a combination of known fragments. In other cases, protein structure prediction has been performed using purely physics-based potentials. A milestone in this field was the work by Duan and Kollman, in which the villin headpiece (a 36-mer) was folded by MD simulations in explicit solvent using a superparallel computer.[6] With the help of worldwide-distributed computing, this small protein was recently also folded by Pande and co-workers[7,8] using implicit and explicit solvent within a total simulation time of 300 and 500 μs, respectively. Another class of physics-based potential energy functions that are successfully used in protein structure prediction are coarse-grained models. A prominent example is UNRES, developed by Scheraga and co-workers.[9] Using this approach, in CASP6, a 102-residue protein was folded to a structure within 7.3 Å of rmsd from its native state.[10]

Despite its great potential, "ab initio" folding can be systematically used only for relatively small proteins, as the computational cost of carrying on an extensive conformational space search is still significant. A possible manner of coping with this problem is to rely on some methodology for accelerating rare events, i.e., conformational changes that involve the crossing of large free energy barriers. A particularly appropriate methodology designed for this scope is replica-exchange molecular dynamics (REMD),[11] in which several replicas of the system are simulated at different temperatures, and exchanges among them are allowed in order to speed up the exploration. This method has been used together with potentials of various accuracies, from coarse-grained to all-atom explicit solvent, to fold small globular proteins.[10,12–16] However, the method requires a great number of replicas, and it works only if the temperature distribution is carefully chosen.[17] This has so far limited the scope of this otherwise extremely powerful methodology. An alternative to the traditional equilibrium approaches is provided by a class of recently developed methods in which the free energy is obtained from nonequilibrium simulations: Wang−Landau sampling,[18] adaptive force bias,[19] and metadynamics.[20] In the latter approach, the dynamics of the system is biased by a history-dependent potential constructed as a sum of Gaussians centered on the trajectory of a selected set of collective variables (CVs). After a transient time, the Gaussian potential compensates the free energy, allowing the system to efficiently explore the space defined by the CVs. This method allows an accurate free energy reconstruction in several variables, but its performance deteriorates with the number of CVs,[20] limiting its usefulness for studying protein folding.

More recently, we introduced a method based on ideas from metadynamics and replica exchange, called bias-exchange metadynamics (BE-META), which is designed specifically for studying complex processes like protein folding.[21,22] In this method, a large set of collective variables is chosen and several metadynamics simulations are performed in parallel, biasing each replica with a time-dependent potential acting on just one or two of the collective variables. Exchanges between the bias potentials in the different variables are periodically allowed according to a replica exchange scheme.[11] Due to the efficaciously multidimensional nature of the bias, the method

10.1021/jp907464b © 2010 American Chemical Society
Published on Web 02/17/2010

allows exploring complex free energy landscapes with great efficiency. Applying this methodology made it possible, with a moderate computational effort, to reversibly fold some small proteins: tryptophane cage, villin, advillin, and insulin.[21–23] The structure of a mutant of advillin that had previously never been investigated was predicted in ref 22, and the numerical prediction was successively validated by NMR experiments. Recently, this methodology was also used to investigate the mechanism of action of cyclophilin[24] and the binding of drugs to proteins.[25,26] These applications demonstrate the potential of BE-META.

The goal of this work is to optimize the performance of this promising technique in order to fold ≈50-residue proteins in the shortest possible time. Both metadynamics and replica exchange, on which BE-META is based, are rather well understood methods, in which convergence and efficiency can be controlled. Instead, BE-META is a more complex approach which requires a setup involving the choice of several free parameters. First of all, the method requires choosing a set of collective variables that guide the exploration of the configuration space. The number (and type) of variables that are used is expected to strongly affect the efficiency of the approach. Moreover, in BE-META, one can choose the dimensionality of the bias on the single replicas, how often the exchange of the bias potentials is attempted, and the height of the Gaussians, which determine how fast the metadynamics bias grows on the single replica. Finally, since in BE-META all the replicas are run at the same temperature, one has to also carefully choose this parameter in order to optimize the performance.

In order to systematically investigate these issues, and to identify general guidelines for optimizing BE-META, one would ideally like to perform several folding simulations for different proteins simulated with diverse models, also with an all-atom description. Unfortunately, this is a very difficult (if not an impossible) task: the computational cost to simulate multiple folding events of a protein in an all-atom force field is just too high. Thus, in order to systematically investigate which BE-META parameters are optimal, we performed extensive tests using a coarse-grained force field, UNRES, which has been successfully used to fold several proteins using replica exchange.[9,10,27] In particular, we considered 1E0G,[28] a 48-residue-long alpha-beta protein that folds with UNRES into its native state.[9] Using this coarse-grained but realistic force field allowed benchmarking the effect of all of the BE-META parameters by repeating several statistically independent simulations. A total of 400 independent folding events were observed during the procedure.

Surprisingly, the parameters that influence the performance of BE-META the most are the temperature and the height of the Gaussians, while the exchange time, the dimensionality of the bias, and the number of collective variables have less influence on the efficiency. We also show that, if the setup is appropriately chosen, BE-META is capable of finding the folded state of 1E0G approximately 10 times faster than normal replica exchange.

II. Methods

A. BE-META. BE-META[21] is a combination of replica exchange[11] and metadynamics,[20] in which multiple metadynamics simulations are performed at the same temperature. Each replica is biased with a time-dependent metadynamics potential acting on one or two different CVs. The bias acting on replica a is defined as

$$V_G^a(x,t) = w\sum_{t'<t} \mathrm{e} - \frac{(s^a(x) - s^a(x(t')))^2}{2\delta s^2} \qquad (1)$$

where x are the coordinates of the system, s^a is the collective variable, and w and δs are the height and width of the Gaussian, respectively. After a certain time (τ_{exch}), exchanges between the different pairs of replicas (a and b) are attempted using the Metropolis scheme[11] with a probability of

$$P_{ab} = \min(1, \exp(\beta[V_G^a(x^a,t) + V_G^b(x^b,t) - V_G^a(x^b,t) - V_G^b(x^a,t)])) \qquad (2)$$

If the move is accepted, the collective variable of replica a performs a jump from $s^a(x^a)$ to $s^a(x^b)$, and the collective variable of replica b, from $s^b(x^b)$ to $s^b(x^a)$. Thus, the trajectory that was previously biased in the direction of the first variable continues its evolution biased by the second (and vice versa). In this manner, a large number of different variables can be simultaneously explored. As in ordinary metadynamics, the Gaussian potential converges to the negative of the free energy. However, the jumps greatly increase the capability of each replica to diffuse through CV space.

B. UNRES Force Field. In UNRES, a polypeptide chain is represented by a sequence of α-carbons, linked by virtual bonds, with attached side chains. United peptide groups are located in the middle of the consecutive α-carbons. Only the peptide groups and the united side chains serve as interaction sites, while the α-carbons assist in the definition of the geometry (see ref 9). All of the virtual bond lengths are held fixed, while the side-chain angles, as well as the virtual-bond angles and the dihedral angles, can vary. UNRES uses physics-based interactions to describe the forces on the coarse-grained particles, and it has been derived as a restricted free energy function of an all-atom polypeptide chain plus the surrounding solvent.

This force field was successful in predicting the native-like structure of several small globular proteins.[27] In CASP6, a 102-residue protein was folded to a structure within 7.3 Å of rmsd from the native state.[10]

C. Implementation of BE-META in UNRES. The molecular dynamics implementation of UNRES[9] provided by A. Liwo in www.chem.univ.gda.pl has been used together with BE-META. The original code has been modified by us, adding to the normal molecular dynamics forces, the forces deriving from the metadynamics potential defined in eq 1:

$$F_{\mathrm{tot}} = F_{\mathrm{unres}} + F_{\mathrm{meta}} \qquad (3)$$

where $F_{\mathrm{meta}} = -\nabla V_G$ (see eq 1). Following the BE-META scheme, a number of metadynamics-UNRES simulations, each biased on one or two different CVs, are run independently. Using an external bash script, exchanges between the biasing potentials are periodically attempted following the Metropolis scheme of eq 2.

D. The Benchmark System: The 1E0G Protein. 1E0G, shown in Figure 1, is a 48 amino acid LysM domain which has a $\alpha\beta\beta\alpha$ secondary structure with the two helices packed at the same side of a two-stranded antiparallel β sheet. This domain, called LysM, was originally identified in enzymes that degrade bacterial cell walls but is also present in many other bacterial proteins. The structure we use here includes residues between 398 and 445,[28] and has been obtained by NMR.

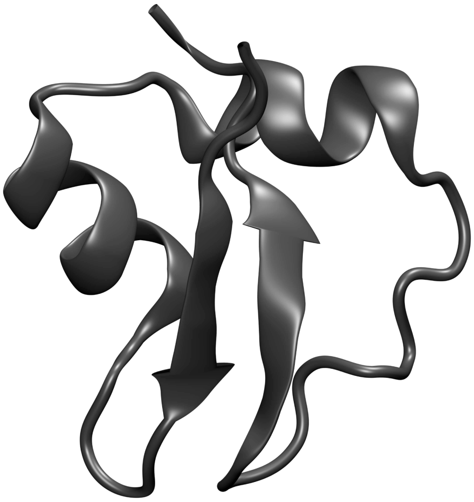

Figure 1. Tertiary structure of protein 1E0G.

This protein has been chosen as a benchmark system because it was folded using the UNRES force field using REMD.[9] To achieve full statistical convergence on the population of the states, the authors used a total simulation time of 80 μs on 80 processors, with 20 temperatures and 16 replicas per temperature. With this level of accuracy, one can be sure that the native fold is the global free energy minimum of the system.

E. The Collective Variables. In refs 21–23, BE-META has been used to fold Trp-cage (PDB code: 1L2Y), villin (1WY3), advillin (1UNC), and insulin chain B (2CEU). The variables used in these studies do not require the *a priori* knowledge of the folded structure but describe essential features of proteins, such as their secondary content and tertiary structure topology. These variables were the number of C_γ and C_α contacts, the number of backbone−backbone H-bonds, the fraction of backbone dihedrals observed in the alpha region of the Ramachandran plot, and the correlation among successive Ψ dihedrals. The CVs that we use in this work are chosen with the scope of testing if the setup in refs 21 and 22 can be improved. Due to the coarse-grained nature of the force field, that does not explicitly describe atoms but only C_α-like and C_β-like coarse-grained particles, the collective variables could not be defined as those in refs 21 and 22. Instead, we redefined contact and dihedral variables in the following coarse-grained manner:

Number of Contacts. This CV counts the number of C_α's in set A that have a distance smaller than 6.5 Å from C_α's in another set B:

$$N = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} C(r_{ij}) \tag{4}$$

where N_A and N_B are the total number of coarse-grained particles in A and in B, respectively, and $C(r_{ij})$ is the switching function:

$$C(r_{ij}) = \frac{1}{1 - (r_{ij}/6.5)^4} \tag{5}$$

and r_{ij} is the distance between the coarse-grained particle i and j measured in angstroms.

Torsional Dihedral Value. It measures the number of dihedrals ϕ_i, in a certain set A, that are similar to a reference dihedral ϕ_o. It is defined as

$$\Phi_{\alpha\beta} = \sum_{i=1}^{N_A} \frac{1 + \cos(\phi_i - \phi_o)}{2} \tag{6}$$

Since we are working in a coarse-grained model, the dihedrals used are those formed between four successive C_α. It was found that for α-helices the appropriate value of ϕ_o is 0.872 rad, while for β-structure $\phi_o = 1.754$ rad. This CV is used to count the number of residues belonging to the α-helix (respectively β-sheet) region of the Ramachandran plot.

Torsional Dihedral Correlation. This CV measures the correlation between two successive dihedrals ϕ_i and ϕ_{i+1} in a certain set A:

$$\Phi_{\mathrm{SS}} = \sum_{i=1}^{N_A} \frac{1 + \cos(\phi_i - \phi_{i+1})}{2} \tag{7}$$

As both in α-helix and β-sheets the dihedrals formed by successive residues have approximately the same value, this CV is able to distinguish if a certain set of dihedrals form secondary structure or not.

Using these variables, we defined the following CVs:

(i) *Number of hydrophobic contacts*: Defined by a variable N (see eq 4) with the set $A = B$ including the coarse-grained particles belonging to hydrophobic residues. In 1E0G, these are the amino acids ILE, LEU, MET, PHE, and VAL. The width of the Gaussian entering in the bias potential (eq 1) is $\delta S = 1.0$.

(ii) *Number of salt bridges*: Defined by a variable N with sets A and B including, respectively, the coarse-grained particles belonging to positively (ARG, HIS, LYS) and negatively (ASP) charged residues in 1E0G ($\delta S = 0.2$).

(iii) *Contacts of the 1st with the 2nd half*: Defined by a variable N with sets A and B including, respectively, the coarse-grained particles belonging to the first and second half of the protein ($\delta S = 2$).

(iv) *Uniformly distributed contacts*: Defined by a variable N with the set $A = B$ including residues 4, 12, 20, 28, 36, and 44 of the protein ($\delta S = 0.2$).

(v) *Torsional dihedral value for the 1st quarter*: Uses $\Phi_{\alpha\beta}$ (see eq 6) with residues 1−12 as set A ($\delta S = 0.2$).

(vi) *Torsional dihedral value for the 2nd quarter*: Uses $\Phi_{\alpha\beta}$ with residues 12−24 as set A ($\delta S = 0.2$).

(vii) *Torsional dihedral value for the 3rd quarter*: Uses $\Phi_{\alpha\beta}$ with residues 24−36 as set A ($\delta S = 0.2$).

(viii) *Torsional dihedral value for the 4th quarter*: Uses $\Phi_{\alpha\beta}$ with residues 36−48 as set A ($\delta S = 0.2$).

(ix) *Torsional dihedral value for the 1st half*: Uses $\Phi_{\alpha\beta}$ with residues 1−24 as set A ($\delta S = 0.2$).

(x) *Torsional dihedral value for the 2nd half*: Uses $\Phi_{\alpha\beta}$ with residues 24−48 as set A ($\delta S = 0.2$).

(xi) *Torsional dihedral correlation for the 1st half*: Uses Φ_{SS} (see eq 7) with residues 1−24 as set A ($\delta S = 0.2$).

(xii) *Torsional dihedral correlation for the 2nd half*: Uses Φ_{SS} with residues 24−48 as set A ($\delta S = 0.2$).

F. Folding Times. In order to assess the performance of BE-META, suitable criteria are needed to compare the efficiency of the different setups. One can either compare the computer time required to reach convergence of the free energy profiles or the computer time required to find the folded state of the protein starting from a completely unstructured state. The first criterion provides more detailed information but has a significant computational cost, as the free energy converges only when the dynamics has explored several times all of the relevant states of the system. In order to reduce the computational cost, and to benchmark the largest possible number of different setups, we have chosen to compare only the average computer time required to find the folded state. It is important to remark that the optimal setup for converging the free energy surface might not be the same as the one for finding in the shortest possible time the folded state.

For each BE-META simulation setup, a statistical analysis was performed on the time required to find the folded state of the protein. As a reference folded state, we used the most populated structure of a REMD run with 6 μs of total simulation time, which has an rmsd from the experimental structure of 5 Å (the same as in ref 9). The protein was considered to be folded when the C_α-rmsd from the reference folded state satisfied the following three conditions: (i) whole-protein rmsd smaller than 5 Å; (ii) β sheet rmsd smaller than 2.5 Å (residues 3−6 and 40−43); (iii) α helices rmsd smaller than 2.5 Å (residues 12−19 and 23−29).

Each simulation was stopped when the three conditions were satisfied in one of the N_R replicas. Denoting by $t_{\text{fold-rep}}$ the time at which this happens, the total simulation time at which the folding event is observed is estimated as

$$t_{\text{fold}} = t_{\text{fold-rep}} N_R \tag{8}$$

Of course, the moment in which the folded state is found is influenced by statistical fluctuations. Thus, in order to assess in a meaningful manner the efficiency of the different setups, $N_{\text{tot}} = 15$ independent runs, initialized with different seeds for the random number generator of the velocities in the Maxwellian distribution, were performed for each setup. As for some setups observing a folding event is very unlikely, we also introduced a maximum simulation time $t_{\max} = N_R \times 100$ ns. The simulation is stopped if this time is reached even if the folded state is not found.

Since in some simulations folding is not observed within $t_{\max}$, the average time required to find the folded state cannot be estimated as the average of the 15 runs. Thus, we fit the observations to an exponential probability distribution $P(t) = e^{-t/\tau_{\text{fold}}}/\tau_{\text{fold}}$ using maximum likelihood.[29] If folding is observed in all of the N_{tot} runs, this approach gives τ_{fold} equal to the average of the observations. Otherwise, τ_{fold} is larger than the average because of the runs in which folding is not observed.

The error bars on τ_{fold} are estimated by using the maximum likelihood procedure.[30] The expression for the error reduces to $\tau_{\text{fold}}/(N_{\text{tot}})^{1/2}$ if folding is observed in all of the N_{tot} runs.

III. Results and Discussion

As our goal is to obtain the optimal BE-META parameters for performing protein folding, several simulations of 1E0G are performed changing selectively:

- the exchange time
- the temperature of the system
- the height of the Gaussians
- the dimensionality of the biasing potential of each replica (one or two).
- the set of CVs

For all of the runs, the same unfolded protein configuration was used as the initial state (rmsd 25 Å from the reference folded state). The integration time step was held fixed for all of the runs, $\Delta t = 5$ fs. The time at which the metadynamics Gaussians are introduced was also held fixed for all of the BE-META simulations, $\tau_G = 1$ ps. The width of the Gaussian (δS) depends on the collective variable (listed in section IIE), and was chosen, like in normal metadynamics, as the (approximate) standard deviation of the value of the collective variable in an unbiased molecular dynamics simulation performed at the folded state.

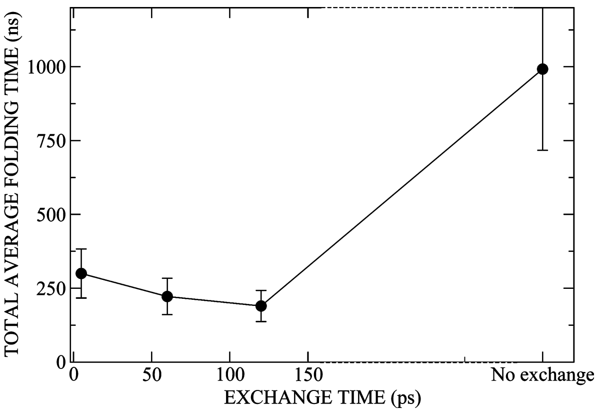

Figure 2. Average folding times for the simulations that use different exchange times.

Using the method described in section IIF, the total average folding time is found for each BE-META setup and they are used to access the efficiency of the different choices for the simulation parameters.

As a reference, the average time required to find the folded state was also computed for a REMD simulation using 20 replicas with temperatures exponentially distributed between 200 and 600 K. Following the procedure described above (section IIF), 15 statistically independent REMD simulations were run, and $\tau_{\text{fold}}^{\text{REMD}} = 380 \pm 98$ ns was found. We will show in the following that, if the BE-META parameters are appropriately chosen, it is possible to observe folding in a significantly shorter time.

A. Does the Exchange Time Influence the Efficiency? As described in section IIA, in BE-META exchanges between the bias potentials in the different variables are attempted every τ_{exch} (exchange time). Normal metadynamics is recovered taking $\tau_{\text{exch}} = \infty$, and the capability of the method of exploring a high dimensional CV space derives from choosing an appropriate exchange frequency (namely, from choosing a finite τ_{exch}). In REMD, it is known that it is marginally better to exchange as often as possible, as exchanges always enhance decorrelation.[31]

The exact role of τ_{exch} in BE-META is less clear: in order to be effective and lead to genuine transitions from different free energy minima, a bias along a specific CV has to be active for the time required to diffuse across the free energy well. In other words, if the direction of the bias changes very frequently (small τ_{exch}), the system could have the tendency to remain close to the free energy minima and not to overcome the barriers. In order to investigate the influence of this parameter, we have performed four BE-META simulations with $\tau_{\text{exch}} = \infty$, 120 ps, 60 ps, and 5 ps. For each simulation, the Gaussian height is held fixed with a value of $w = 0.012$ kcal/mol, $T = 280$ K, and the set of CVs i−viii with eight one-dimensional replicas is used. In Figure 2, the total average folding times for these simulations are shown.

From these results, it is found that very frequent or very rare exchanges make BE-META marginally less efficient. The optimal exchange time for this system is 120 ps; this is sufficient to give the system enough time to explore the local free energy wells before the direction of the bias changes. It is clear from the error bars that the effect of any finite τ_{exch} on the efficiency

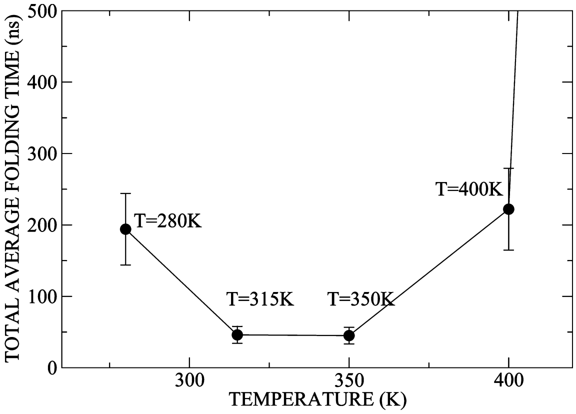

Figure 3. Average folding times for BE-META simulations at different temperatures.

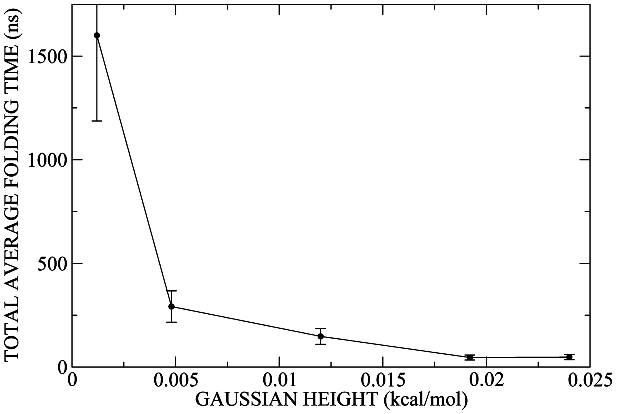

Figure 4. Average folding times for the BE-META simulations that use different Gaussian heights.

is rather small if compared to other parameters. However, if the exchanges are not performed at all ($\tau_{\text{exch}} = \infty$), folding is observed within an average time that is ~5 times larger. This shows that using BE-META (with practically any finite τ_{exch}) leads to a significant improvement of the capability of the system in exploring the configuration space. It is somehow surprising that, for this system, it is possible to explore folding in a reasonably short time biasing a single variable by normal metadynamics ($\tau_{\text{exch}} = \infty$). To the best of our knowledge, if an all-atom force field is used, observing folding under these conditions would be very unlikely. This might indicate a tendency of UNRES to overstabilize the folded state or to the barriers.

B. What is the Optimal Temperature for Performing BE-META? In normal REMD, the replicas span a wide temperature range and their number increases with the number of degrees of freedom of the system.[17] Instead, in BE-META, all of the replicas are run at the same T. This is an advantage if the protein is modeled with an explicit solvent Hamiltonian, since introducing the solvent does not require an increase of the number of replicas.

Five BE-META simulations at 280, 315, 350, 400, and 500 K are performed with the aim of finding the optimal temperature. All of the runs use $\tau_{\text{exch}} = 20$ ps, $w = 0.012$ kcal/mol, and a set of four CVs: i, iv, ix, and x. The average folding times for these simulations are shown in Figure 3.

As the temperature increases, the system is able to find faster the folded state of the protein, with the best performance between 315 and 350 K. Comparing this with the results reported by Liwo in ref 9 for the specific heat of this system, it is found that the temperature at which BE-META has its best performance is also approximately where the maximum of the specific heat is located. Since the transition probability is proportional to $e^{-\Delta E/k_BT}$, it is reasonable that the temperature where exploration of the space is optimal will be close to the critical temperature (T_F). For $T > T_F$, the system will mostly explore unfolded structures, and thus will not efficiently localize the folded state of the system. At very high T ($\geq$500 K for this case), the folded state is not explored at all.

C. What is the Effect of the Gaussian Height? In metadynamics, the dynamics is biased by a history-dependent potential (eq 1). The larger the height of the Gaussian (w), the larger the extra forces, the more the system is out of equilibrium, and the larger the error on the reconstructed free energy surface.[20]

In the case of BE-META, the role of w is investigated by running simulations with different Gaussian heights: $w =$ 0.0012, 0.0048, 0.012, 0.0192, and 0.024 kcal/mol. The runs are performed with $\tau_{\text{exch}} = 20$ ps at a temperature of 280 K and with eight one-dimensional replicas acting on the 8 CVs i−viii. The average folding times as a function of w are shown in Figure 4.

From the picture, it is clear that the efficiency in finding the folded state increases significantly by raising the height of the Gaussian. As the Gaussians get higher, the biasing forces increase and the system is encouraged to move faster through the conformational space. Still, there is a limit to the height of the Gaussians, because if the biasing forces are much larger than those of the force field, the simulation will eventually collapse. For the system considered here, this happens if a w larger than ~0.3 kcal/mol is used. We remark that the optimal setup for finding efficiently the folded state can be different from the optimal setup for estimating accurately the free energy of all the conformations.

D. One or Two-Dimensional Gaussians? The efficiency of normal metadynamics is reduced when the dimensionality of the Gaussians increases, as it takes more time to fill a high dimensional space. However, if the free energy of a system is intrinsically high dimensional and one uses only a one-dimensional bias, the metadynamics estimate of the free energy is affected by large errors.[32] BE-META provides, at least in principle, a solution to this problem, as one can explore high dimensional free energy landscapes by using low dimensional biases. Still, in BE-META, one can choose one- or two-dimensional biases on the single replicas. We here investigate this issue comparing three BE-META setups, all using $\tau_{\text{exch}} =$ 20 ps, $w = 0.12$ kcal/mol, $T = 280$ K, and the same set of eight CVs i−viii. The three setups differ only in the dimensionality of the Gaussians (one or two) and the number of replicas used:

- one-dimensional Gaussians and eight replicas
- four replicas with two-dimensional biases: Rep. 1, i−ii; Rep. 2, iii−iv; Rep. 3, v−vi; Rep 4, vii−viii
- eight replicas with “overlapping” two-dimensional biases: Rep. 1, i−ii; Rep. 2, ii−iii; Rep. 3, iii−iv; Rep. 4, iv−v; Rep 5, v−vi; Rep 6, vi−vii; Rep 7, vii−viii; Rep 8: viii−ix.

In Figure 5, the total average folding times are shown for these simulation setups.

From the picture, it is evident that using lower dimensional Gaussians increases the efficiency of the method by a factor of ~2. The difference in performance might be due to the fact that for the two-dimensional case it is necessary to put more Gaussians in order to make the system explore new regions of the free energy landscape. For all three setups, we verified that

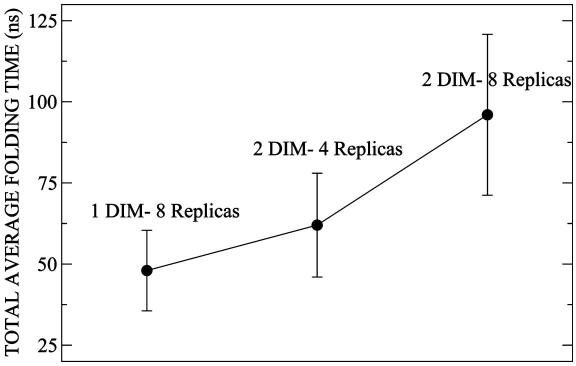

Figure 5. Average folding times for the simulations that use one or two-dimensional hills with different numbers of replicas.

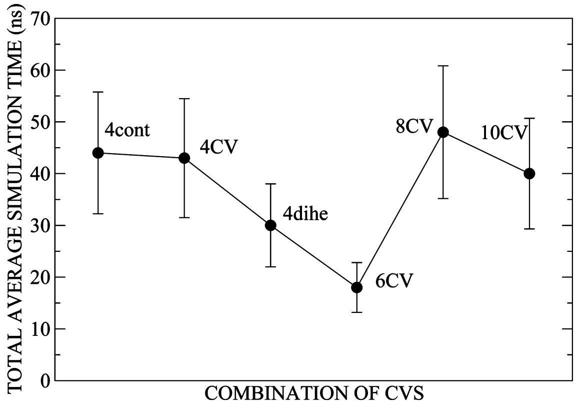

Figure 6. Average folding times for the BE-META simulations that use different sets of CVs.

using a larger w does not improve significantly the performance (we are in the plateau region of Figure 4). Setups 2 and 3 use two-dimensional Gaussians, but one has four replicas, the other eight. The setup with more replicas is marginally less efficient, since for a large number of replicas, including replicas which bias irrelevant variables, the average folding time becomes larger due to its definition (eq 8) even if folding is observed at the same time on the single replica.

E. What Kind of Variables Are Essential for Localizing Efficiently the Folded State? An essential issue in all enhanced-sampling techniques based on CVs is their proper choice. If a crucial CV is forgotten, these methods do not work. We here consider the CVs described in section IIE, which are a coarse-grained version of those that have been successfully used for protein folding in previous works.[21,22] These variables measure the content of secondary and tertiary structure of the protein, and are not based on the *a priori* knowledge of the folded state. In order to understand how many variables and which CVs are essential to observe folding in this system, six different CV combinations, of the variables described in section IIE, are attempted:

- 10 CVs: i−viii and xi−xii (4 contacts − 4 torsional dihedrals − 2 correlation dihedrals)
- 8 CVs: i−viii (4 contacts − 4 torsional dihedrals)
- 6 CVs: i−iv, ix, and x (4 contacts − 2 torsional dihedrals)
- 4 CVs: i, iv, ix, and x (2 contacts − 2 torsional dihedrals)
- 4 contacts: i−iv
- 4 torsional dihedrals: v−viii

Considering the results of the previous sections, we have decided to use here the optimal parameteres found before: each BE-META simulation is run with $\tau_{\mathrm{exch}} = 20$ ps, $T = 315$ K, $w = 0.12$ kcal/mol, and one-dimensional Gaussians. The total average folding time is found for each different CV setup, using the procedure described in section IIF. The results are shown in Figure 6.

All of the simulation setups, including the ones with only four variables, are able to find the folded state of 1E0G within ~50 ns of total simulation time. We can conclude that, rather surprisingly, the choice of the CV setup does not influence dramatically the efficiency of the method. Within the error bars, all of the setups lead to an approximately similar efficiency.

The only choice that leads to marginally better results is the six CV setup. This might be rationalized assuming that the two following effects influence the efficiency: if a large number of replicas N_R is used, including replicas which bias irrelevant degrees of freedom, the folding time increases linearly with N_R (see eq 8). Instead, for a too small number of replicas, not all of the relevant degrees of freedom are sufficiently biased. However, it is clear from Figure 6 that these effects are small for protein 1E0G. This is an encouraging result, because it shows that the method is not very sensitive to the choice and number of CVs, provided the most relevant degrees of freedom are biased.

It cannot be excluded that this particularly favorable situation might be due to the force field, that could overstabilize the folded state with respect to other misfolded structures, or lower the barriers.

In other words, even if the variables used in this work are sufficient to fold 1E0G with UNRES, this might not be the case if an all-atom force field with explicit solvent is used.

IV. Conclusions

In this work, we have carefully benchmarked bias-exchange metadynamics (BE-META) with the scope of optimizing its performance in protein folding. To this aim, we performed several BE-META simulations of the 1E0G protein, that is known to fold to its native structure with the UNRES force field. The parameters entering in BE-META have been systematically varied in order to find the setup that allows localizing the folded state in the shortest possible time. More than 400 independent folding events were observed in a total computational time of 100 μs. The results of this analysis are the following:

- Using BE-META with practically any finite exchange time τ_{exch} leads to a significant improvement in the capability of the system of exploring the configuration space with respect to normal metadynamics. Very frequent or very few exchanges make BE-META marginally less efficient: the optimal τ_{exch} should give the system enough time to diffuse through the local free energy wells but also the opportunity to experience changes in the direction of the bias.
- BE-META's best performance in finding the folded state of the protein is obtained when the system is run as close as possible to its folding temperature (T_F). This allows a fast exploration of the degrees of freedom not explicitly included in the bias. However, for $T > T_F$, the system will mostly explore unfolded structures, and thus will not efficiently localize the folded state of the system.
- The height of the Gaussian influences significantly the efficiency of BE-META in finding the folded state. A relatively large w should be used. Indeed, the forces coming from the time-dependent bias bring the system out of equilibrium, but they allow a fast exploration of the conformational space. Still, after a certain value of w, the performance does not improve and reaches a plateau (see Figure 3). Moreover, there is a limit to the height of the

Gaussians that can be used: if the biasing forces become much larger than those of the force field, the simulation will eventually explode.

- The dimensionality of the Gaussians used in BE-META influences the efficiency of the method only marginally. A slightly better performance is obtained if one-dimensional Gaussians are used. This is possibly due to the fact that for the two-dimensional case it is necessary to put more Gaussians in order to make the system explore new regions of the free energy landscape.
- The choice and number of the CVs that are biased, provided that the most relevant degrees of freedom are included, do not influence dramatically the time required to localize the folded state. Marginally better results can be obtained if the CV setup has a small number of replicas.

Overall, the parameters that have a larger influence on BE-META's efficiency are the temperature at which the system is run and the height of the biasing Gaussian. If the Gaussian height is too low, or the temperature is far from T_F, the exploration of the conformational space becomes much slower.

When using the optimal setup (CVs i, ii, iii, iv, ix, and x with six one-dimensional replicas, $T = 315-370$ K, $w > 0.02$ kcal/mol, $\tau_{\mathrm{exch}} = 20-120$ ps), it is possible to fold systematically 1E0G within ~20 ns of total simulation time. The time required to fold the same system with replica exchange (REMD) is more than an order of magnitude higher (~390 ns). Even if the BE-META setup is not optimal, the performance of the method remains adequate (~100 ns for most of the setups). This makes the approach a viable candidate for attempting to fold average-size proteins by an accurate all-atom explicit solvent force field.

It should be remarked that, in this work, we have only considered one single protein modeled with a coarse-grained force field. One could argue that the optimal choice of the BE-META setup might possibly depend on the specific protein and force field. Thus, future research for optimizing the methodology with different proteins modeled by accurate force fields is still required. Instead, REMD retains the advantage of its simplicity and generality, as it does not require choosing system-specific collective variables, or system-dependent parameters.

Acknowledgment. We are grateful to Stefano Piana for several useful suggestions, and to CINECA for providing computational resources.

References and Notes

(1) Zhang, Y. Progress and challenges in protein structure prediction. *Curr. Opin. Struct. Biol.* **2008**, *18*, 342–348.

(2) Monticelli, E.; Sorin, E.; Tieleman, D.; Pande, V.; Colombo, J. Molecular simulation of multistate peptide dynamics: A comparison between microsecond timescale sampling and multiple shorter trajectories. *J. Comput. Chem.* **2008**, *29*, 1740–1752.

(3) Jones, D.; Taylor, W.; Thornton, J. A new approach to protein fold recognition. *Nature* **1992**, *358*, 86–89.

(4) Bradley, P.; Misura, K.; Baker, D. Toward high-resolution de novo structure pre-diction for small proteins. *Nature* **2005**, *309*, 1868–1871.

(5) Wu, S.; Skolnick, J.; Zhang, Y. Ab initio modeling of small proteins. *BMC Biol.* **2007**, *5*, 7.

(6) Duan, Y.; Kollman, P. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **1998**, *282*, 740–744.

(7) Zagrovic, B.; Snow, C.; Shirts, M.; Pande, V. Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *J. Mol. Biol.* **2002**, *323*, 927–937.

(8) Jayachandran, G.; Vishal, V.; Pande, V. Using massively parallel simulation and Markovian models to study protein folding: Examining the dynamics of the villin headpiece. *J. Chem. Phys.* **2006**, *124*, 164902.

(9) Liwo, A.; Khalili, M.; Czaplewski, C.; Kalinowski, S.; Oldziej, S.; Wachucik, K.; Scher-aga, H. Modification and Optimization of the United-Residue (UNRES) Potential En-ergy Function for Canonical Simulations. I. Temperature Dependence of the Effective Energy Function and Tests of the Optimization Method with Single Training Proteins. *J. Phys. Chem. B* **2007**, *111*, 260–285.

(10) Oldziej, S.; Czaplewski, C.; Liwo, A.; Chinchio, M.; Nanias, M.; Vila, J.; Khalili, M.; Arnau-tova, Y.; Jagielska, A.; Makowski, M. Physics based protein-structure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7547–7552.

(11) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141.

(12) Zhou, R. Trp-cage: Folding free energy landscape in explicit water. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13280–13285.

(13) Chebaro, Y.; Dong, X.; Laghaei, R.; Derreumaux, P.; Mousseau, N. Replica Ex-change Molecular Dynamics Simulations of Coarse-grained Proteins in Implicit Solvent. *J. Phys. Chem. B* **2009**, *113*, 267–274.

(14) Yoda, T.; Sugita, Y.; Okamoto, Y. Cooperative Folding Mechanism of a beta-Hairpin Peptide Studied by a Multicanonical Replica-Exchange Molecular Dynamics Simula-tion. *Proteins* **2007**, *66*, 846–859.

(15) Lei, H.; Wu, C.; Liu, H.; Duan, Y. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 4925–4930.

(16) Paschek, D.; Nymeyer, H.; Garcia, A. Replica exchange simulation of reversible folding/unfolding of the Trp-cage miniprotein in explicit solvent: On the structure and possible role of internal water. *J. Struct. Biol.* **2006**, *157*, 542–533.

(17) Trebst, S.; Troyer, M.; Hansmann, U. Optimized parallel tempering simulations of proteins. *J. Chem. Phys.* **2006**, *124*, 174903.

(18) Wang, F.; Landau, D. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* **2001**, *86*, 2050.

(19) Darve, E.; Pohorille, A. Calculating free energies using average force. *J. Chem. Phys.* **2001**, *115*, 9169–9183.

(20) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562–12566.

(21) Piana, S.; Laio, A. A bias-exchange approach to protein folding. *J. Phys. Chem. B* **2007**, *111*, 4553–4559.

(22) Piana, S.; Laio, A.; Marinelli, F.; Van Troys, M.; Bourry, D.; Ampe, C.; Martins, J. Predicting the effect of a point mutation on a protein fold: The villin and advillin headpieces and their Pro62Ala mutants. *J. Mol. Biol.* **2008**, *375*, 460–470.

(23) Todorova, N.; Marinelli, F.; Piana, S.; Yarovsky, I. Exploring the Folding Free Energy Landscape of Insulin Using Bias Exchange Metadynamics. *J. Phys. Chem. B* **2009**, *113*, 3556–3564.

(24) Leone, V.; Lattanzi, G.; Molteni, C.; Carloni, P. Mechanism of Action of Cy-clophilin A Explored by Metadynamics Simulations. *PLoS Comput. Biol.* **2009**, *5*, e1000309.

(25) Pietrucci, F.; Marinelli, F.; Carloni, P.; Laio, A. Substrate Binding Mechanism of HIV-1 Protease from Explicit-Solvent Atomistic Simulations. *J. Am. Chem. Soc.* **2009**, *131*, 11811–11818.

(26) Kranjc, A.; Bongarzone, S.; Rossetti, G.; Biarnes, X.; Cavalli, A.; Bolognesi, M.; Roberti, M.; Legname, G.; Carloni, P. Docking Ligands on Protein Surfaces: The Case Study of Prion Protein. *J. Chem. Theory Comput.* **2009**, *5*, 2565–2573.

(27) Liwo, A.; Czaplewski, C. Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. *J. Chem. Phys.* **2001**, *115*, 2323–2343.

(28) Bateman, A.; Bycroft, M. The structure of a LysM domain from E. coli membrane-bound lytic murein transglycosylase (MltD). *J. Mol. Biol.* **2000**, *229*, 1113–1119.

(29) Kay, S. M. *Fundamentals of Statistical Signal Processing*; Estimation Theory: Prentice Hall, NJ, 1993; Vol. 221, pp 1−67.

(30) Dempster, A.; Laird, N.; Rubin, D. Maximum likelihood from incomplete data via em algorithm. *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–38.

(31) Zhang, W.; Wu, C.; Duan, Y. Convergence of replica exchange molecular dynamics. *J. Chem. Phys.* **2005**, *123*, 154105.

(32) Laio, A.; Gervasio, F. L. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep. Prog. Phys.* **2008**, *71*, 126601.

JP907464B