JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Free Energy Guided Sampling

Ting Zhou and Amedeo Caflisch*

Department of Biochemistry, University of Zurich, CH-8057 Zurich, Switzerland

Supporting Information

ABSTRACT: A free energy-guided sampling (FEGS) method is proposed for accelerating exploration of conformational space in unbiased molecular dynamics. Using the cut-based free energy profile and Markov state models, FEGS speeds up sampling of the canonical ensemble by iteratively restarting multiple short simulations in parallel from regions of the free energy surface visited rarely. This *exploration* stage is followed by a *refinement* stage in which multiple independent runs are initiated from Boltzmann distributed conformations. Notably, FEGS does not require either collective variables or reaction coordinates and can control the kinetic distance from the starting conformation. We applied FEGS to the alanine dipeptide, which has a human-comprehensible two-dimensional free energy landscape, and a three-stranded antiparallel β-sheet peptide of 20 residues whose folding/unfolding process is governed by a delicate interplay of enthalpy and entropy. For these two systems, FEGS speeds up the exploration of conformational space by 1 to 2 orders of magnitude with respect to conventional sampling and preserves the basins and barriers on the free energy profile.

INTRODUCTION

Molecular dynamics (MD) simulations are a powerful tool for studying (macro)molecular structure and flexibility, as they generate atomistic details of system evolution. In principle, if MD simulations are run long enough, all of the relevant states are visited and the population converges to the (global) equilibrium. In practice, computer hardware and the complexity of the system limit the simulation time. Even though the reversible folding of structural peptides and small proteins has been investigated by millisecond MD simulations with explicit water[1] and implicit solvent,[2] inadequate sampling is still the largest source of errors in MD simulations.

Sampling can be hindered by deep enthalpic basins on the free energy surface,[3] and long simulation time will be consumed for the system to escape from these enthalpic traps. The excessive sampling in these traps does not help to understand the transitions among states, because those transitions are usually slow and associated with high energy barriers out of those traps. Several methods have been developed to deal with the sampling problem. In one class of methods, e.g., replica exchange MD (REMD),[4] high temperatures are used to cross the enthalpy barriers, low temperatures are employed to sample the detail of the free energy landscape, and a random walk in temperature space maintains canonical sampling at each temperature. Recently kinetic network models have been introduced to extract kinetics from REMD sampling[5] and to optimize the parameters of REMD.[6] However, the entropic part of the temperature-dependent transition rate ($k_a(T) = k_0 \exp(\Delta S/k_B) \exp(-\Delta H/k_B T)$) is unaltered by increasing the temperature.[7] In practice, if the goal is the single-temperature sampling, its efficiency will be immediately decreased by a factor corresponding to the number of replicas, which is usually large for systems with explicit solvent.[8−10] Therefore, it is very difficult for REMD to gain more than an order of magnitude speedup at physiological temperature.[7]

Another class of methods is noncanonical sampling where a bias is introduced to increase the possibility of slow conformational transitions. The bias often depends on a finite number of predetermined degrees of freedom (e.g., collective variables in metadynamics and path-optimization approaches) that can describe transitions of interest.[11−15] In those methods, the choice of variables is nontrivial, since only the immediate vicinity of the path determined by collective variables can be meaningfully investigated. Nonetheless, in the glassy region of the free energy surface, the transitions are involved in a large variety of paths with similar energy profiles,[16] which, in theory, demand the use of a large amount of collective variables. In practice, the calculation increases exponentially with the amount of collective variables, because the biases have to be added to the potential function. Recently, Tribello et al. developed reconnaissance metadynamics, which allows one to bias the free energy profile efficiently with a very large number of low-dimensional and locally valid collective variables.[17] However, collective variables and biases still have to be implemented in MD simulation codes.

The Markov State Model (MSM) is an elegant and useful tool to investigate systems that undergo (large-scale/long-time conformational) transitions.[18−21] By using the MSM, the weight of each state in equilibrium MD can be estimated by long-term distribution of each state in the MSM. The global equilibration is not required anymore. Moreover, the kinetic distance can be estimated by the mean first passage time (mfpt).[19] On the basis of MSMs, Huang and co-workers have developed the adaptive seeding method (ASM) for sampling the folding/unfolding dynamics of biological systems, which shows that MSMs can be used to recover the correct equilibrium populations from nonequilibrium simulations.[22] In ASM, high temperatures are used to flatten the enthalpy traps and obtain the broad sampling; therefore, the sampling range is difficult to control. The method is appropriate for complete folding/unfolding processes (e.g., eight-nucleotide RNA hairpin 5′-GCUUUUGC-3′ shown in ref 22) but is expected to be less adequate for local movement, e.g., protein conformational

Received: February 20, 2012
Published: April 23, 2012

   dx.doi.org/10.1021/ct300147t | J. Chem. Theory Comput. 2012, 8, 2134−2140

transitions[23,24] and ligand binding/unbinding.[25−28] In general, it is difficult to control the range of sampling in methods that make use of enhanced simulation temperature.

In this paper, we propose a method for constant-temperature, unbiased, reaction-coordinate-free, and range-controlled MD sampling guided by the barrier-preserving, cut-based free energy profile (cFEP)[29,30] determined on the fly. The method is termed free energy guided sampling (FEGS). We show that FEGS does not require sampling at unwanted temperatures, oversampling at the enthalpic traps, or the choice of collective variables. Moreover, the sampling can be restricted within the kinetic range of interest. Importantly, the convergence of the free energy profile[30,31] can be used to indicate the sufficiency of sampling.

To evaluate the efficiency and the accuracy of FEGS, two systems are investigated: (1) The first is the alanine dipeptide, which can be *completely* sampled and whose free energy landscape can be projected onto the human-comprehensible Ramachandran map (ϕ,ψ-map).[32] For the alanine dipeptide, the free energy landscape produced by FEGS is shown to converge much faster than by conventional sampling (CS). (2) The second is a 20-residue β-sheet peptide (Beta3s),[33,34] which folds reversibly to its correct NMR structure with an efficient implicit solvent model,[35] and whose folding process can be described as a delicate balance of enthalpy and entropy.[31] For these two systems, FEGS visits clusters of conformations and generates a converged cFEP much more efficiently than CS.

METHODOLOGY

FEGS. The sampling can start from a single conformation or an ensemble of structures (e.g., an energy minimized crystal structure or NMR conformer bundle). For efficiency, two iterative sampling stages are used: the exploring stage and the refining stage (Figures 1 and 2).

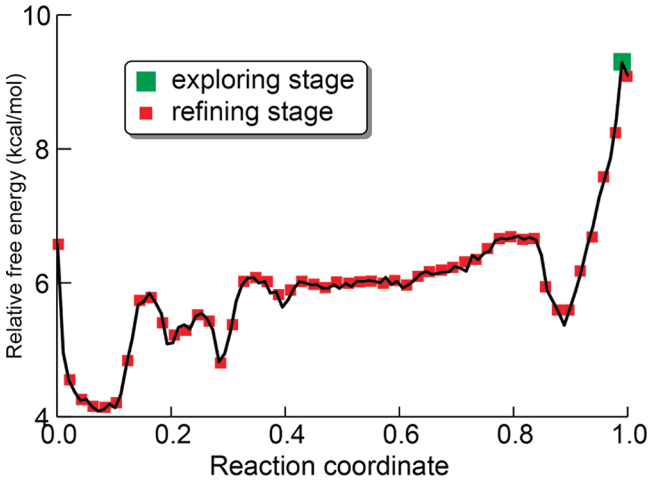

Figure 1. Schematic illustration of the selection of restarting conformations in FEGS. The black curve is the cFEP calculated from the sampling data of current and previous iterations. If the FEGS is in the exploring stage, new MD runs start from the conformation indicated by the green square with different initial velocities. In the refining stage, new MD runs start from conformations indicated by red squares.

In the exploring stage, the system is driven kinetically as far as possible from the starting conformation. The distance is measured with mfpt calculated using the MSM. The general strategy is as follows: (e1) start n_{expl} simulations from the initial conformation; (e2) cluster snapshots into mesostates; (e3) build an MSM and

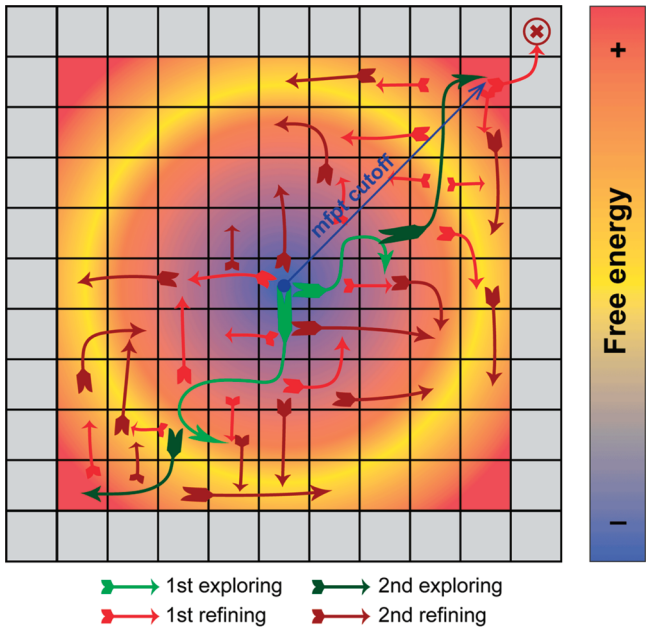

Figure 2. Schematic illustration of two-stage sampling of free energy landscape by FEGS. The starting structure is shown here in the center of the grid (blue dot). In the first exploring stage, the sampling moves away from the starting structure (light green arrows). Then, in the second exploring stage (dark green arrows), the MD runs start from the furthest conformation obtained previously. After reaching the mfpt cutoff (blue arrow), the first and second steps of refining (light and dark red arrows, respectively) start from conformations according to their Boltzmann probabilities. Gray squares are the conformational space beyond the mfpt cutoff. No further sampling will start from these gray squares, even though they have been visited by an exploring/refining sampling (red "×" at top right corner).

calculate the equilibrium population of each mesostate; (e4) sort mesostates by mfpt to the starting conformation; (e5) calculate cFEP; (e6) restart n_{expl} simulations from the barriers farthest to the starting conformation (the green square in Figure 1); (e7) continue with step e2 until the maximum of mfpt reaches the cutoff, or the maximal time of exploration is reached. The free energy barriers can be overcome in the exploring stage because the simulations are started from the kinetically farthest regions. In the canonical ensemble, the system rarely visits those regions, and as a consequence the transitions over high barriers are extremely rare events.[12,36] However, most of the mesostates (i.e., details of the free energy landscape) between the starting state and the farthest state are overlooked in the exploring stage.

In the refining stage, overlooked mesostates are sampled extensively. The choice of restarting conformations for the next iteration of sampling is as follows: instead of selecting the farthest barriers, n_{ref} MD runs are initiated from conformations that are equally distributed along the cFEP (red squares in Figure 1). The refining stage stops when the cFEP converges, which can be checked automatically by measuring the covariance of the latest free energy profile and the previous one. The sampling can also re-enter the exploring stage during or after the refining stage if the mfpt value of the kinetically farthest mesostate is smaller than the cutoff. This re-entering could happen because, as more sampling data are used for building the MSM, the mfpt becomes more accurate. By re-starting sampling iteratively, the system will not be trapped into the deep free energy basins (Figure S1 in the Supporting Information). The simulation time for sampling will be equally

dx.doi.org/10.1021/ct300147t | J. Chem. Theory Comput. 2012, 8, 2134−2140

distributed on the conformational space. The algorithm is summarized in the flowchart in Figure 3.

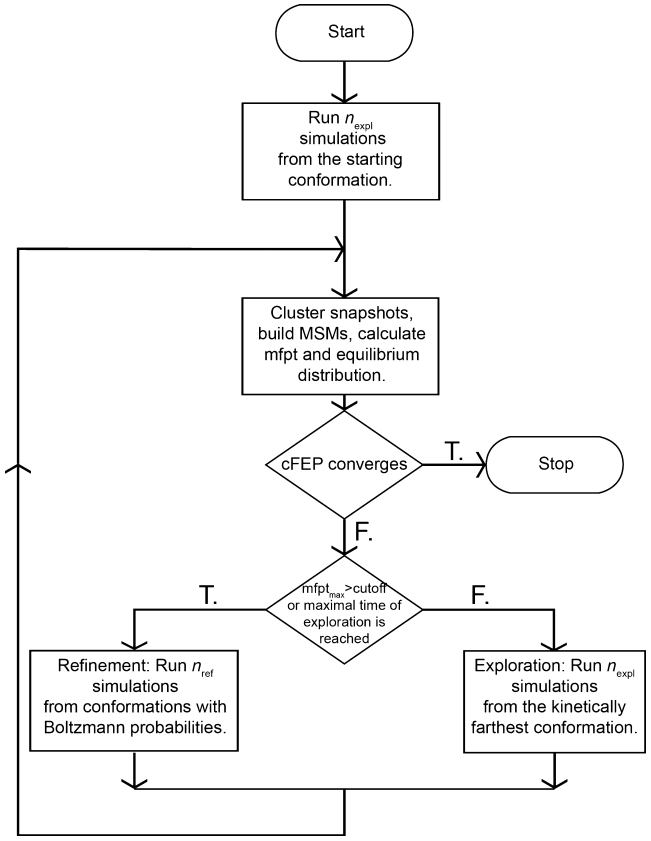

Figure 3. Flowchart of the FEGS method. FEGS requires only a starting structure and a value of $mfpt_{max}$.

For the simulations of local motion, e.g., loop conformational transition and ligand binding/unbinding, the kinetic range of sampling can be controlled with mfpt. The system quickly jumps across free energy barriers and reaches the farthest conformation within interesting mfpt range in the exploring stage. To efficiently push the system far away from the starting conformation, short simulations are used in the exploring stage because, like CS, long simulations can be easily "trapped" by enthalpic basins and hence cannot effectively explore the free energy landscape. Restarting sampling frequently at the barrier furthest from the initial conformation gives a greater chance to reach the unvisited regions than continuous samplings. After the border of interesting mfpt is reached, the refining stage starts (red arrows in Figure 2). At each refining stage, n_{ref} simulations are initiated from conformations with their Boltzmann probabilities, which are calculated from the previous trajectories. All of the information from previous samplings is taken into account for guiding the samplings toward the equilibrium. Note that interim Boltzmann probabilities are not necessarily accurate but will eventually converge to the accurate one as more mesostates are discovered,[21] and the precise transition probabilities among them are iteratively approached. That is, the free energy profile does not change much between successive iterations of refining.

Markovian State Models from Multiple MD Trajectories. The time scale of *in silico* simulations is usually much shorter than the time scale of biological interest. MSMs are a powerful tool used to predict long-term properties (both population of each state and kinetics) of a system that cannot be simulated adequately with present computers. To build MSMs, the simulated system needs to be locally equilibrated.[22] In FEGS, multiple independent constant temperature MD runs are initiated from a single conformation. Thus, the transition network[29,31,37] is intrinsically connected, and MSMs can be straightforwardly built from this network. Formally, the transition network extracted from MD trajectories is directed. That is, let n_{ij} be the absolute number of transitions from mesostate i to mesostate j; n_{ij} and n_{ji} are not necessarily equal. However, in a closed, isolated, and classical system, the detailed balance ($n_{ij} = n_{ji}$) always holds,[38] and the directed graph can be simplified to an undirected graph. In practice, due to incomplete sampling, it is quite often the case that the transition network satisfies neither detailed balance nor ergodicity.[39] The arithmetic average of n_{ij} and n_{ji} is commonly used as the number of transitions between mesostates i and j in the undirected graph.[40] Nevertheless, this averaging imposes the long-term equilibrium which is adequate only for extensive sampling very close to it.[39] In FEGS, the following approach is used to generate an irreducible graph from a reducible graph without discarding any sampled conformations. In a reducible graph, the transition network is weakly connected (i.e., there are pairs of mesostates i and j, for which $n_{ij} > 0$ but $n_{ji} = 0$). If forward transitions ($i \rightarrow j$) are observed, but no backward transition ($i \rightarrow j$), a single backward transition, which is physically the observable minimum, is added $n_{ji} = 1$ (Figure S2 in the Supporting Information). By this postprocessing, the transition network becomes strongly connected or ergodic, and the system being studied can be formalized as a finite, homogeneous, and regular Markov chain. The transition probabilities can be calculated as $T_{ij} = n_{ij}/\sum_k n_{ik}$, where T is the transition (probability) matrix. The steady state of the Markov chain Π (π_1, π_2, ..., π_n) can be determined by solving the system of linear equations $\Pi = \Pi T$, where π_i is the steady (long-term) distribution of mesostate i.[41] Utilizing all visited mesostates, this postprocessing approach makes FEGS efficient in exploring the conformational space. Note that the statistical error introduced by adding a backward transition will lead the simulation to start from unsampled regions and will be corrected in the following iterations of the FEGS protocol. The aforementioned approach is only used for building interim MSMs to guide samplings. These MSMs do not necessarily represent the quantitatively accurate kinetics of the system. Finally, an accurate free energy profile can be attained using the largest ergodic component identified by Tarjan's algorithm to preserve the statistics of the original network.[39,42]

RESULTS AND DISCUSSION

Alanine Dipeptide. To illustrate the efficiency of FEGS, we compared the evolution of the free energy surface calculated by CS and FEGS (Figure 4). At the beginning of the CS, the system remained trapped mainly in the region with $\phi < 0°$. Strikingly, after 5 ns, FEGS visited all of the four free energy minima. Since 100 ns, the free energy surface sampled by FEGS had essentially converged, while the one sampled by CS had not yet visited the α_L region.

The one-dimensional cFEP can be used to quantitatively analyze the transition network of the alanine dipeptide.[29] The height of the barriers and the population of the basins as sampled by CS and FEGS are essentially identical (Figure S3 in Supporting Information).

To evaluate the speedup of FEGS with respect to CS, we plotted the ratio of the time needed for the system to visit a

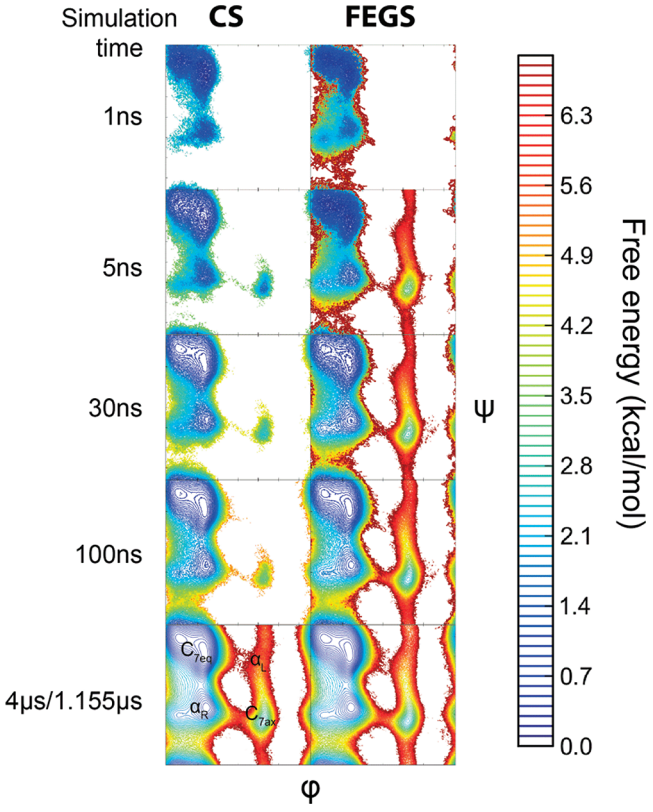

Figure 4. Comparison of CS (left) and FEGS (right) of the alanine dipeptide. The contour plots show that the CS is slower than FEGS in leaving the α-helical (termed α_R, $\phi < 0°$ and $-60° < \psi < 0°$) and β-strand (termed C_{7eq}, $\phi < 0°$ and $0° < \psi < 180°$) regions. The regions of C_{7ax} and α_L ($\phi > 0°$) are sampled much more efficiently by FEGS than CS. A movie clip (Mov S1) that shows the efficiency of FEGS is included in the Supporting Information. Each iteration in the exploring stage of FEGS consists of 10 independent MD runs of 10 ps each, while each iteration in the refining stage consists of 1000 MD runs of 20 ps each. The exploring and the refining stages contain 50 and 57 iterations, respectively.

certain number of clusters defined as bins of size 1° on both ϕ and ψ angles of the Ramachandran map. The speedup of FEGS did not reach 10-fold until the first 60 000 clusters were visited (Figure S4 in the Supporting Information) because at the beginning of the sampling both methods were trapped in the C_{7eq} region, where the simulations were started. In FEGS, the system escaped this first basin when most of the clusters in it were visited. After sampling about 75 000 clusters, the system sampled by FEGS escaped the $\phi < 0°$ half of the map and maximized the speedup. The unsupervised learning protocol of FEGS helps the system to find the positions of restart samplings for quickly escaping the present basin. The more data the transition network contains, the more accurate restarting positions it will suggest, and thus the larger the possibility of escaping the basin. This is the reason that in the early stage of sampling, the speedup of FEGS is not significant. Once any one of the independent MD runs of a FEGS iteration succeeds to escape the basin spontaneously, others will start from the region out of the basin in the next iteration of FEGS. The speedup decreases from its maximum 77-fold to about 50-fold as the sampling is elongated because it becomes more and more difficult to find an unvisited cluster in the relatively narrow conformational space of the alanine dipeptide.

Beta3s Peptide. Folding of the 20-residue β-sheet peptide Beta3s (whose sequence in one-letter code is TWIQNGSTK-WYQNGSTKIYT) has been extensively investigated by MD simulations.[5,30,31,34,40,43−45] Using CS and an implicit solvent model,[35] Beta3s folds reversibly from its heterogeneous denatured state to its native structure, a three-stranded antiparallel β-sheet.[33]

The FEGS protocol consisted of 50 iterations of exploration containing 100 independent simulations of 100 ps each, and 12 iterations in the refining stage containing 100 independent 20-ns simulations each. In the exploring stage and the beginning of refining stage, the cFEP fluctuated because of insufficient statistics (Figure S5 in the Supporting Information). In the exploring stage, the covariance between cFEPs obtained at consecutive iterations showed a large variability ranging from −0.06 to 0.37. The cFEP converged, as more trajectories were iteratively appended for calculating the transition matrix. After the 55th iteration, the covariances were always larger than 0.42 (Figure S6 in the Supporting Information).

FEGS is about 1 order of magnitude faster in sampling of mesostates and transitions for building a converged MSM than CS (Figure S7 in the Supporting Information). Without being trapped in enthalpic basins, compared to CS, FEGS needed less CPU time to visit a similar amount of mesostates and to observe a similar number of transitions for Beta3s (Table 1).

Table 1. Comparison of Transition Matrix between FEGS and CS

sampling type	FEGS	CS[a]		
length of simulation (μs)	24.5	22	22	44
number of mesostates	321349	171402	180582	311674
number of nonzero elements in transition matrix	2194315	970548	1057461	2033897
ratio[b]	6.83	5.66	5.86	6.53

[a]Ten independent CSs were initiated from the NMR native conformation. Every sampling generated a 4.4 μs trajectory. For evaluating convergence, the 10 trajectories were separated into two groups, each of which contained five trajectories (22 μs in total). [b]The ratio of the number of nonzero elements in the transition matrix and the number of the mesostates.

Moreover, FEGS has good reproducibility. To check the convergence, five FEGS runs of 24.5 μs each were started from the native structure of Beta3s. In those five FEGS runs (with different random seeds for assigning initial velocities in the first exploring iteration), the standard deviation of the populations of the native basin and the helical basin, which is kinetically farthest from the native conformation, are 1.1% and 2.9%, respectively (Figure 5). The standard deviation of the heights of the free energy barriers that separate these two basins from the rest of the free energy surface are 0.07 and 0.09 kcal/mol, respectively (Figure S8 in the Supporting Information). The secondary structural annotation of each mesostate shows that cFEP can cluster trajectories generated by both FEGS and CS into kinetically relevant states. From the native basin, the kinetic ordering of states was similar in both algorithms, but the CS stayed longer than the FEGS in the native basin, where the sampling started. Thus, very large mesostates in the native basin were observed in the cFEP obtained by CS (Figure 5). Due to the dependence of native basin size on the simulation length in CS (Figure S9 in the Supporting Information), very long runs have to be carried out to attenuate the effects of the identical, rather than Boltzmann

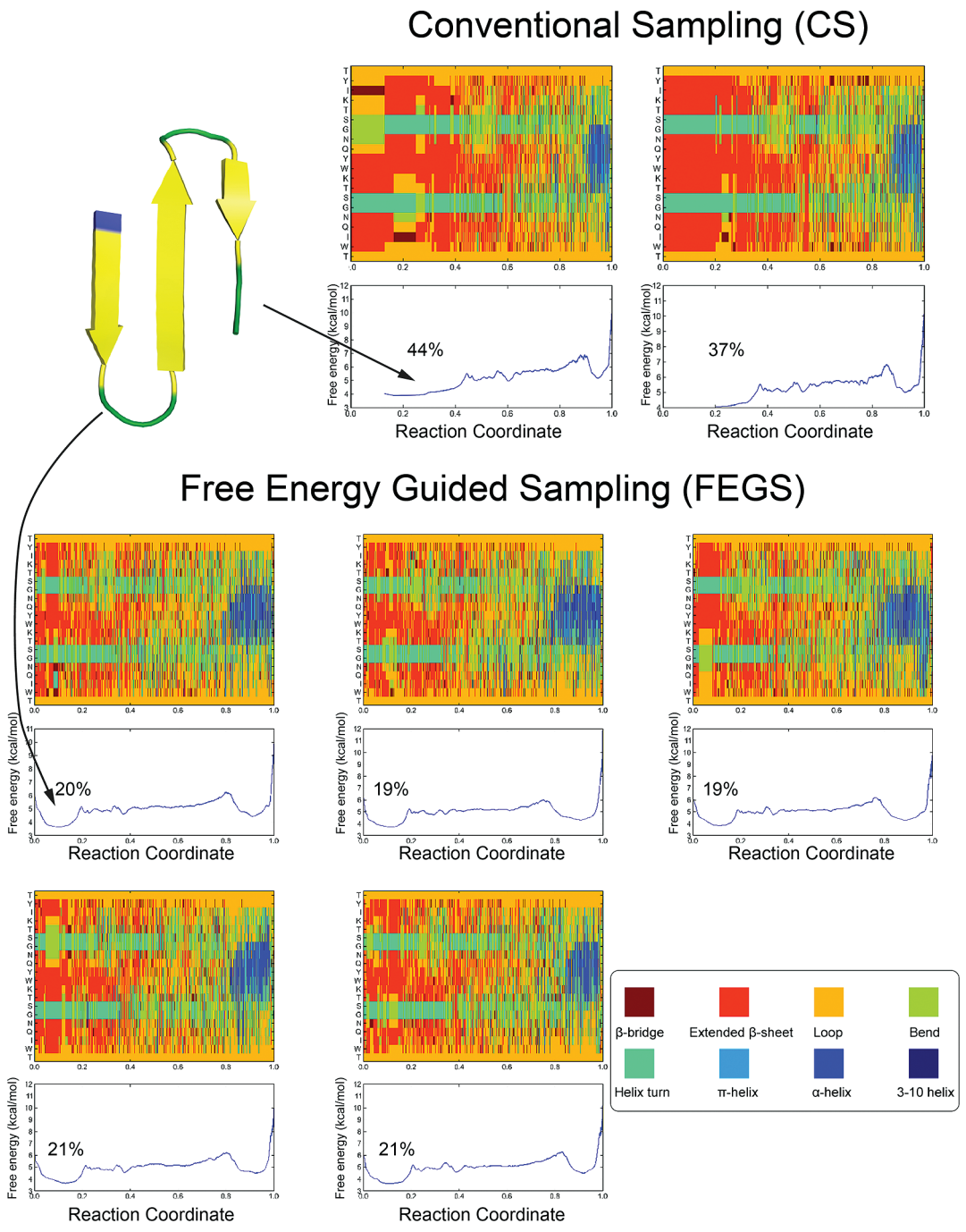

Figure 5. Comparison of CS and FEGS of Beta3s. Both CS and FEGS started from the native structure (top left). All cFEPs are plotted using the native state as a reference. The population of the native basin, which is the basin on the left up to the first energy barrier, is denoted in percentage on each cFEP. (Top) Each of the two cFEPs was calculated using 22 μs of CS obtained by five independent MD runs of 4.4 μs each. (Middle and bottom) Each of the five cFEPs was calculated using 24.5 μs of FEGS. The upper part of each panel (with the sequence of the Beta3s on the y axis) shows the colored DSSP[52] strings of the cluster representatives, which are arranged according to the reaction coordinate of the cFEP. The legend of colors for different secondary structure elements in the traces is indicated in the bottom right panel. FEGS shows a remarkable convergence of the cFEP. In contrast, CS does not seem to have converged (see also Figures S8 and S9 in the Supporting Information).

distributed, starting conformations.[39,46] In contrast, FEGS quickly loses the bias due to a single starting structure because it iteratively restarts from the MSM estimate of the Boltzmann distribution.

■ CONCLUSIONS

We have introduced the FEGS method which uses the cut-based free energy profile and Markov state model to efficiently explore the conformational space of peptides and proteins by conventional MD. The efficiency and accuracy of FEGS was demonstrated by applications to the alanine dipeptide and reversible folding of a 20-residue β-sheet peptide. The former illustrates in a human-comprehensible manner its speedup efficacy, as the two-dimensional free energy surface converges faster by more than 1 order of magnitude for FEGS than CS. The latter peptide is a challenging system because of the

complex denatured state consisting of enthalpic traps and an entropically stabilized non-native helical basin. FEGS generates a converged cFEP of Beta3s by sampling mesostates about 1 order of magnitude faster than CS. The two applications presented here used implicit solvent simulations of peptides because the free energy surface can been fully characterized. Yet extension to explicit water simulations is straightforward, and we are currently using FEGS to sample the conformational space of the catalytic domain of a tyrosine kinase which will be reported elsewhere.

The FEGS method is a bias-free and reaction-coordinate-free approach. FEGS samples heuristically, and no additional information is needed except for a starting conformation. Since FEGS does not need to bias the energy function, the simulation code, force field parameters, and/or coarse-grained models do not need to be modified. It samples at the temperature of interest and does not visit regions of conformational space populated only at elevated temperatures. The sampling controlled by mfpt also makes possible investigating local movements, such as ligand binding/unbinding and conformational transitions involving localized structural elements. The FEGS protocol uses MD simulation programs solely to generate trajectories. More precisely, it provides initial conformations for the iterative restarting of simulations and does not modify the kernel of the simulation codes (e.g., energy functions, integration algorithms). Thus, it can straightforwardly be employed with the majority of MD programs. Moreover, FEGS could be used with Monte Carlo methods by taking into account that the mfpt would not be a physical time but an effective kinetic distance. Finally, FEGS can be accomplished in an embarrassingly parallel way and is suitable for running on modern computer clusters.

APPENDIX

Simulation Setup

The MD simulation protocols of CS and FEGS are identical. For the alanine dipeptide, Langevin dynamics with a friction coefficient equal to 50 ps^{-1} was used. The trajectories at 300 K were generated with the CHARMM[47] program using polar hydrogen energy function PARAM19 and saved every integration step (2 fs). The SHAKE algorithm was applied to hydrogen atoms.[48] The effective solvation free energy was approximated with the SASA implicit solvation model.[34] The reference cFEP and the ϕ,ψ map under the global equilibrium were generated using a 4-μs trajectory (a total of 2×10^9 conformations were analyzed).

The MD simulations of the three-stranded β-sheet peptide Beta3s were performed using the Leapfrog algorithm implemented in CHARMM with the PARAM19 and the SASA solvation model. The temperature was controlled with the Berendsen thermostat (coupling every 5 ps) at 330 K. By applying the SHAKE algorithm, a time step of 2 fs was used, and the trajectories were saved every 0.2 ps for a total of about 1×10^8 snapshots.

Simulation Detail

The alanine dipeptide snapshots were binned using 1° resolution on both ϕ and ψ angles of the Ramachandran map. For Beta3s, the WORDOM[49] implementation of the sequential leader-like clustering algorithm was used with a threshold of 2.5 Å on the pairwise coordinate root mean square deviation of the unsymmetrical heavy atoms of residues 3 to 18.

Mean First Passage Time (mfpt)

Given the transition matrix T of the MSM, the mfpt of mesostate i to the reference mesostate A is the solution of the linear equations $\mathrm{mfpt}_i = \Delta t + \sum_j (T_{ij} \times \mathrm{mfpt}_j)$ with an initial boundary condition $\mathrm{mfpt}_A = 0$,[50] where Δt corresponds to the lag time used for building the MSM.

Cut-Based Free Energy Profile (cFEP)

The input for the cFEP calculation is the network of conformational transitions, which is derived from the direct transitions between clusterized snapshots (mesostates of the network) sampled at a given time interval (2 fs for alanine dipeptide and 0.2 ps for Beta3s) along the MD simulations. For each mesostate, mesostates are partitioned into two groups using the values of the mfpts to the reference mesostate to define a cut. The free energy is related to the flow across the cut and approximated as $\Delta G = -kT \ln Z_{AB}$ where Z_{AB} is the partition function of the mfpt-based cutting surface (for further details, see ref 29−31). The result is a one-dimensional profile along a reaction coordinate (the relative partition function) that preserves the barrier height between well-separated free energy basins.

Markovianity Test

The Markovian property is not critical for the interim MSMs, as the sampling itself is not the final one. Yet the Markovianity of the models built after each iteration was evaluated by calculation of the non-Markovian flux,[19] which is a variant of the Chapman−Kolmogorov test.[51] The non-Markovian flux of the MSMs (with lag time equal to the saving interval) was always below 5% during the exploring stage, and in the last refining iteration it was 0.057% and 2.0% for the alanine dipeptide and Beta3s, respectively.

ASSOCIATED CONTENT

Supporting Information

Comparison of distribution of the cluster size of FEGS and CS on Beta3S (Figure S1). Schematic illustration of manually added backward transition for generating an ergodic transition network (Figure S2). cFEPs of the alanine dipeptide calculated from the transition networks sampled by CS and FEGS (Figure S3). The ratio of total simulation time needed by CS and FEGS to visit a certain number of clusters of the alanine dipeptide (Figure S4). Evolution of the cut-based free energy profile of Beta3s (Figure S5). The covariances between two adjacent cFEPs of Beta3s (Figure S6). The ratio of total simulation time needed by CS and FEGS to visit a certain number of clusters of the 20-residue peptide Beta3s (Figure S7). cFEPs of Beta3s in the reproducibility test which consisted of five independent FEGS runs of 24.5 μs each started from the native structure (Figure S8). Comparison of cFEPs generated by CS of Beta3s with different trajectory length (Figure S9). Comparison of the sampling speed of CS and FEGS for the alanine dipeptide (Mov S1). This material is available free of charge via the Internet at http://pubs.acs.org/.

AUTHOR INFORMATION

Corresponding Author

*Phone: (+41 44) 635 55 21. Fax: (+41 44) 635 68 62. E-mail: caflisch@bioc.uzh.ch.

Funding Sources

This work was supported by a grant of the Swiss National Science Foundation to A.C.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Riccardo Scalco for useful discussions and Dr. Andreas Vitalis for the code for plotting the secondary structure of each cluster. We also thank Christian Bolliger for the management of the Schrödinger cluster of the University of Zurich, which was used for MD simulations and data processing.

■ DEDICATION

Dedicated to Martin Karplus whose passion for biology, chemistry, and physics has strongly inspired the research activities of A.C. during the past 20 years.

■ ABBREVIATIONS

MSM, Markov State Model; MD, molecular dynamics; mfpt, mean first passage time; cFEP, cut-based free energy profile; FEGS, free energy guided sampling; CS, conventional sampling; ASM, adaptive seeding method; RMSD, root-mean-square deviation; REMD, replica exchange MD

■ REFERENCES

(1) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517.
(2) Settanni, G.; Rao, F.; Caflisch, A. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 628.
(3) Luo, G.; Andricioaei, I.; Xie, X. S.; Karplus, M. *J. Phys. Chem. B* **2006**, *110*, 9363.
(4) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141.
(5) Muff, S.; Caflisch, A. *J. Phys. Chem. B* **2009**, *113*, 3218.
(6) Zheng, W.; Andrec, M.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 15340.
(7) Zuckerman, D. M.; Lyman, E. *J. Chem. Theory Comput.* **2006**, *2*, 12001202.
(8) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 13749.
(9) Okur, A.; Wickstrom, L.; Layten, M.; Geney, R.; Song, K.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2006**, *2*, 420.
(10) Kamberaj, H.; van der Vaart, A. *J. Chem. Phys.* **2007**, *127*, 234102.
(11) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562.
(12) Faradjian, A. K.; Elber, R. *J. Chem. Phys.* **2004**, *120*, 10880.
(13) Maragliano, L.; Vanden-Eijnden, E. *Chem. Phys. Lett.* **2006**, *426*, 168.
(14) Laio, A.; Gervasio, F. L. *Rep. Prog. Phys.* **2008**, *71*, 126601.
(15) Pan, A. C.; Sezer, D.; Roux, B. *J. Phys. Chem. B* **2008**, *112*, 3432.
(16) Czerminski, R.; Elber, R. *J. Chem. Phys.* **1990**, *92*, 5580.
(17) Tribello, G. A.; Ceriotti, M.; Parrinello, M. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 17509.
(18) Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571.
(19) Guarnera, E.; Pellarin, R.; Caflisch, A. *Biophys. J.* **2009**, *97*, 1737.
(20) Pande, V. S.; Beauchamp, K.; Bowman, G. R. *Methods* **2010**, *52*, 99.
(21) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. *J. Chem. Phys.* **2011**, *134*, 174105.
(22) Huang, X.; Bowman, G. R.; Bacallado, S.; Pande, V. S. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19765.
(23) Gan, W.; Yang, S.; Roux, B. *Biophys. J.* **2009**, *97*, L8.
(24) Berteotti, A.; Cavalli, A.; Branduardi, D.; Gervasio, F. L.; Recanatini, M.; Parrinello, M. *J. Am. Chem. Soc.* **2009**, *131*, 244.
(25) Shan, Y.; Seeliger, M. A.; Eastwood, M. P.; Frank, F.; Xu, H.; Jensen, M. Ø.; Dror, R. O.; Kuriyan, J.; Shaw, D. E. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 139.
(26) Huang, D.; Caflisch, A. *PLoS Comput. Biol.* **2011**, *7*, e1002002.
(27) Buch, I.; Giorgino, T.; Fabritiis, G. D. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 10184.
(28) Huang, D.; Caflisch, A. *ChemMedChem* **2011**, *6*, 1578.
(29) Krivov, S. V.; Karplus, M. *J. Phys. Chem. B* **2006**, *110*, 12689.
(30) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 13841.
(31) Krivov, S. V.; Muff, S.; Caflisch, A.; Karplus, M. *J. Phys. Chem. B* **2008**, *112*, 8701.
(32) Hermans, J. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 3095.
(33) de Alba, E.; Santoro, J.; Rico, M.; Jiménez, M. A. *Protein Sci.* **1999**, *8*, 854.
(34) Ferrara, P.; Caflisch, A. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 10780.
(35) Ferrara, P.; Apostolakis, J.; Caflisch, A. *Proteins* **2002**, *46*, 24.
(36) Vanden-Eijnden, E.; Venturoli, M. *J. Chem. Phys.* **2009**, *130*, 194101.
(37) Noé, F.; Fischer, S. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154.
(38) Van Kampen, N. G. *Stochastic Processes in Physics and Chemistry*, 2nd edition; North-Holland Personal Library: North Holland, 1992, Vol. 1.
(39) Scalco, R.; Caflisch, A. *J. Phys. Chem. B* **2011**, *115*, 6358.
(40) Muff, S.; Caflisch, A. *J. Chem. Phys.* **2009**, *130*, 125104.
(41) Kemeny, J. G.; Snell, J. L. *Finite Markov Chains: With a New Appendix "Generalization of a Fundamental Matrix" (Undergraduate Texts in Mathematics)*; Springer: New York, 1976.
(42) Tarjan, R. *SIAM J. Comput.* **1972**, *1*, 146.
(43) Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299.
(44) Muff, S.; Caflisch, A. *Proteins* **2008**, *70*, 1185.
(45) Qi, B.; Muff, S.; Caflisch, A.; Dinner, A. R. *J. Phys. Chem. B* **2010**, *114*, 6979.
(46) Smith, L. J.; Daura, X.; van Gunsteren, W. F. *Proteins* **2002**, *48*, 487.
(47) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545.
(48) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.
(49) Seeber, M.; Felline, A.; Raimondi, F.; Muff, S.; Friedman, R.; Rao, F.; Caflisch, A.; Fanelli, F. *J. Comput. Chem.* **2011**, *32*, 1183.
(50) Apaydin, M. S.; Brutlag, D. L.; Guestrin, C.; Hsu, D.; Latombe, J.-C.; Varma, C. *J. Comput. Biol.* **2003**, *10*, 257.
(51) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19011.
(52) Andersen, C. A. F.; Palmer, A. G.; Brunak, S.; Rost, B. *Structure* **2002**, *10*, 175.

■ NOTE ADDED AFTER ISSUE PUBLICATION

The caption to Figure 1 was cut off in the PDF version of this paper published on the web on April 23, 2012 and in the June 2012 issue. The caption is now corrected in the PDF version published August 13, 2012, and an Addition and Correction appeared on the web on August 13, 2012, and in issue 9.