JCTC Journal of Chemical Theory and Computation

Article

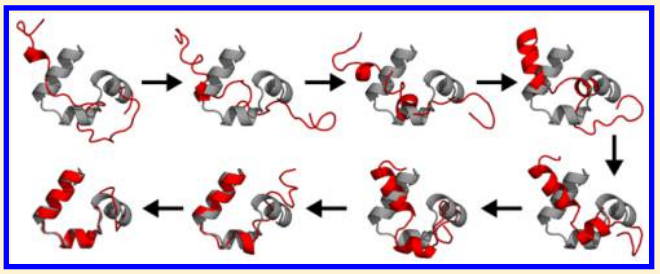
pubs.acs.org/JCTC

FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs

Maxwell I. Zimmerman[†] and Gregory R. Bowman*[,†,‡,§]

[†]Department of Biochemistry & Molecular Biophysics, [‡]Department of Biomedical Engineering, and [§]Center for Biological Systems Engineering, Washington University School of Medicine, St. Louis, Missouri 63110, United States

ABSTRACT: Molecular dynamics simulations are a powerful means of understanding conformational changes. However, it is still difficult to simulate biologically relevant time scales without the use of specialized supercomputers. Here, we introduce a goal-oriented sampling method, called fluctuation amplification of specific traits (FAST), for extending the capabilities of commodity hardware. This algorithm rapidly searches conformational space for structures with desired properties by balancing trade-offs between focused searches around promising solutions (exploitation) and trying novel solutions (exploration). FAST was inspired by the hypothesis that many physical properties have an overall gradient in conformational space, akin to the energetic gradients that are known to guide proteins to their folded states. For example, we expect that transitioning from a conformation with a small solvent-accessible surface area to one with a large surface area will require passing through a series of conformations with steadily increasing surface areas. We demonstrate that such gradients are common through retrospective analysis of existing Markov state models (MSMs). Then we design the FAST algorithm to exploit these gradients to find structures with desired properties by (1) recognizing and amplifying structural fluctuations along gradients that optimize a selected physical property whenever possible, (2) overcoming barriers that interrupt these overall gradients, and (3) rerouting to discover alternative paths when faced with insurmountable barriers. To test FAST, we compare its performance to other methods for three common types of problems: (1) identifying unexpected binding pockets, (2) discovering the preferred paths between specific structures, and (3) folding proteins. Our conservative estimate is that FAST outperforms conventional simulations and an adaptive sampling algorithm by at least an order of magnitude. Furthermore, FAST yields both the proper thermodynamics and kinetics, allowing for a direct connection with kinetic experiments that is impossible with many other advanced sampling algorithms because they provide only thermodynamic information. Therefore, we expect FAST to be of great utility for a wide range of applications.



INTRODUCTION

Understanding the structural mechanisms of conformational changes, such as protein folding and allosteric communication, is a notoriously difficult problem. Molecular dynamics (MD) simulations can complement experimental studies of such problems by filling in information beyond their reach, such as an atomically detailed picture of conformational heterogeneity. However, it is extremely difficult to simulate biologically relevant processes on millisecond and slower time scales with conventional molecular dynamics simulations.

Three broad classes of methods have been developed to capture longer time scale processes with computer simulations. The first class consists of directed methods that actively drive simulations toward some goal, such as steered molecular dynamics,[1] metadynamics,[2,3] the string method,[4,5] and methods for introducing restraints from experiments.[6,7] Unfortunately, these often go through unrealistically high-energy conformations (Figure 1, red path) and fail to explore conformations orthogonal to the direction in which they are being driven, though new methods are more capable of finding the energetically preferred paths.[8] The second class consists of undirected methods that attempt to accelerate the exploration of all conformations, such as replica exchange,[9] accelerated molecular dynamics,[10] weighted ensembles,[11−13] combinations of coarse-grained and all-atom simulations,[14] and adaptive sampling.[15−21] While these methods will eventually provide the correct result, conformational space is so enormous that researchers can easily expend all of their computing resources exploring structures that are not relevant to the problem they set out to solve (Figure 1, yellow enclosed space). Most of the approaches in these two classes also preclude the acquisition of kinetic information by introducing a biasing force or altering properties like the potential energy or temperature. While they still provide the proper thermodynamics, the lack of kinetic information makes it impossible to make quantitative connections with many experimental techniques. The third class of methods focuses on the development of a specialized supercomputer, such as a distributed computing platform[22,23] or purpose-built hardware,[24] that is capable of running enough simulation to discover the relevant conformational space. This approach has led to some of the most dramatic demonstrations

Received: August 3, 2015
Published: November 11, 2015

DOI: 10.1021/acs.jctc.5b00737
J. Chem. Theory Comput. 2015, 11, 5747−5757

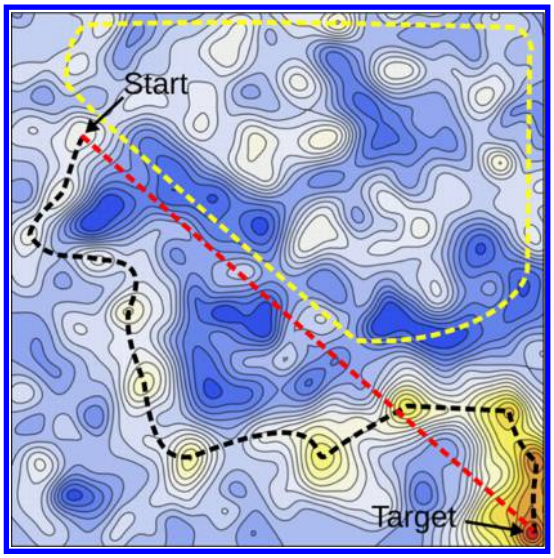


Figure 1. Contour plot of an energy landscape colored in blue, white, yellow, and red from highest to lowest energy. The black line is the optimal path from a starting state to a target. The red line is the path found by directed methods. The yellow line encompasses the area where undirected methods are likely to get lost.

of the power of simulations, including insights into protein folding[25,26] and allosteric communication[27−29] on up to millisecond time scales. However, there are still many processes beyond the reach of these computers. Moreover, very few researchers have access to these resources.

Here, we propose a goal-oriented sampling method called fluctuation amplification of specific traits (FAST) that combines elements of directed and undirected searches to quickly explore regions of conformational space that are relevant to a given problem. This algorithm was inspired by the fact that a protein folds by following an energy gradient to its native state,[30−32] but following such gradients is nontrivial because there are energy barriers and dead ends along the way. We hypothesized that the correlation between structures and energies gives rise to similar gradients for many other physical properties, such as the root-mean-squared deviation (RMSD) to a target structure and the solvent-accessible surface area. For example, we expect that transitioning from a conformation with a small solvent-accessible surface area to one with a large surface area will require passing through a series of conformations with steadily increasing surface areas. If these gradients exist, then it should be possible to follow them to identify structures that maximize (or minimize) specific physical properties. Literature on optimization theory has dealt with related problems by balancing trade-offs between focused searches around promising solutions (exploitation) and trying completely novel solutions (exploration).[33,34] The FAST algorithm leverages these ideas (1) to recognize and amplify structural fluctuations along gradients that optimize a selected physical property whenever possible, (2) to overcome barriers that interrupt these overall gradients, and (3) to reroute to discover alternative paths when faced with insurmountable barriers.

FAST achieves these objectives by drawing on work on the multi-armed bandit problem and particle-swarm optimization. The multi-armed bandit problem[33] is a classic exploration/exploitation trade-off problem in which a hypothetical gambler at a row of slot machines must decide when to (1) try a relatively untested slot machine that could easily yield enormous or meager returns and (2) when to exploit the expected rewards of a tried-and-true machine. A key result is that one can obtain outstanding performance by using estimates of the uncertainty in the expected rewards for each slot machine to select the one that has the highest probability of yielding the greatest rewards.[35] A simple means of achieving this objective is to always choose the slot machine with the highest probability of the greatest return, which can be assessed by a reward function of the form

$$r(i) = \mu(i) + \alpha\sigma(i) \quad (1)$$

where i is a slot machine, $\mu(i)$ is its average return, $\sigma(i)$ is the standard deviation of the returns from that machine, and α is a constant that controls the importance of uncertainty.[36] Particle-swarm optimization[34] is another means of addressing exploration/exploitation trade-offs, but it does so by using a swarm of walkers to explore parameter space. These walkers are designed to balance between spreading out to explore different potential solutions and converging on promising regions of parameter space.

Inspired by these ideas, FAST runs successive swarms of simulations where the starting points for each swarm are chosen from the set of all previously discovered conformations based on a reward function. This reward function quantifies the relative likelihood that simulations started from different structures will discover new conformations that maximize (or minimize) a selected physical property. It mimics the functional form of eq 1 by including a directed component that parallels the mean return and an undirected component corresponding to the uncertainty in the return on investment, as described in the Methods section. The directed component allows FAST to follow gradients by searching near promising solutions for even better ones. Following such gradients alone is not an ideal search strategy because some regions of conformational space with a promising gradient may lead to dead ends. To avoid this pitfall, the undirected component favors poorly sampled regions of conformational space, allowing the algorithm to recognize dead ends where simulations repeatedly fail to discover structures that better optimize the target function and to reroute to less explored regions of conformational space in search of new leads. Since no biasing force is applied to any individual simulation, the final data set can be used to build a Markov state model (MSM) to extract the proper thermodynamics and kinetics despite the nonequilibrium distribution of starting points for the trajectories (see the Methods section for details).[37−39] This approach differs from existing adaptive sampling techniques[15−19] in that it seeks to prioritize what types of structures are explored rather than purely trying to minimize the statistical uncertainty in a model. This is an important distinction because adaptive sampling can easily exhaust finite computational resources searching through irrelevant conformations, whereas we expect the goal-oriented method presented here to quickly focus in on regions of conformational space that are relevant to the problem at hand.

To test FAST, we have applied it to three challenging sampling problems: (1) the discovery of unexpected pockets that might be valuable drug targets, (2) the identification of transition paths between specific conformations, and (3) protein folding. We begin by retrospectively analyzing existing MSMs to assess whether various physical properties have the gradients we hypothesize to exist in protein conformational

DOI: 10.1021/acs.jctc.5b00737
J. Chem. Theory Comput. 2015, 11, 5747−5757

space. Then, we test FAST's ability to identify and follow gradients that are relevant to each of the problems considered.

METHODS

FAST Algorithm. The FAST algorithm is intended to optimize any selected geometric function ϕ of a protein structure, including, but not limited to, energies, RMSDs, and solvent-accessible surface areas. For a given physical property ϕ, the FAST−ϕ algorithm is

(1) Start a swarm of simulations from a set of initial conformations, such as one or more known crystal structures.
(2) Cluster all the simulation data collected so far into discrete conformational states.
(3) Calculate a reward function for each state

$$r_\phi(i) = \bar{\phi}(i) + \alpha\bar{\psi}(i) \quad (2)$$

where i is a particular state, $\bar{\phi}(i)$ is a directed component that fosters exploitation by favoring states that optimize some structural metric of interest (such as the RMSD to a target) compared to other states, $\bar{\psi}(i)$ is an undirected component that fosters exploration by favoring states that are poorly sampled compared to other states, and α is a control parameter that determines the relative importance of the directed and undirected components of the reward function. The bars over each component of this reward function indicate that we feature-scale them (equations below) to highlight the differences between states and ensure that a variable with a greater dynamic range does not overshadow the other component. For example, when trying to maximize the solvent-accessible surface area, $\bar{\phi}(i)$ will range from zero for the state with the lowest solvent-accessible surface area to one for the state with the largest solvent-accessible surface area and $\bar{\psi}(i)$ will range from zero for the most sampled state to 1 for the least sampled state. Therefore, poorly sampled states that optimize the target function are expected to yield the highest reward, while states that have been explored thoroughly and are far from the target are not expected to be rewarding.
(4) Start a new swarm of simulations, where the number of simulations started from each state is proportional to the reward function for that state.
(5) Repeat steps 2−4 until the target function has converged or until some predetermined amount of simulation has been conducted.
(6) Build an MSM from the final data set to capture the proper thermodynamics and kinetics, thereby correcting for any bias introduced by selecting starting conformations for each swarm of simulations according to our reward function instead of a Boltzmann distribution.[37,38]

It is important to note that a valid MSM does not need to be constructed for each round of FAST. This is an important feature since the algorithm needs to work properly even when there is not enough data to accurately estimate transition probabilities for parts of conformational space. The clustering simply needs to be at a resolution that is fine-grained enough to distinguish (1) structures with different values of the target geometric function and (2) regions of conformational space that are well-sampled versus those that are poorly sampled. In step 6, more care is required to build a valid MSM that satisfies the Markov assumption, has a reasonable lag time, and captures the phenomena of interest.

Feature-scaling transforms some quantity into a ranking that ranges from 0 to 1 from the least preferred to the most preferred value, respectively. For a quantity ϕ that one wishes to maximize

$$\bar{\phi}(i) = \frac{\phi(i) - \phi_{\min}}{\phi_{\max} - \phi_{\min}}$$

whereas for a quantity one wishes to minimize

$$\bar{\phi}(i) = \frac{\phi_{\max} - \phi(i)}{\phi_{\max} - \phi_{\min}}$$

where $\phi_{\min}$ and $\phi_{\max}$ are the minimum and maximum values of $\phi(i)$, respectively.

For the undirected component of our reward function, $\bar{\psi}(i)$, we adopt a Bayesian perspective to devise a simple measure of how likely simulations started from a given state are to discover new states. We begin by assuming that the biomolecule under consideration has n structural states and that $n = n_d + n_u$, where n_d is the number of states FAST has discovered so far and n_u is the number of undiscovered states. Following previous work,[15,16] we assume that, prior to observing any data, a simulation started from some initial state has an equal probability of transitioning to any possible final state. Formally, this is achieved by adding a pseudocount $\tilde{C} = 1/n$ to every element of a transition count matrix (C) used to keep track of the number of transitions observed between every pair of states (C_{ij} is the number of transitions observed from state i to state j). Next, we assume that the transition probabilities out of each state are Dirichlet-distributed, which is a common way to enforce that they are properly normalized.[15,40,41] Given this assumption, the expected probability of transitioning from state i to any undiscovered state in the set u is

$$E(p_{iu}) = \sum_{j\in u}\left[\frac{1+\tilde{C}}{\sum_{k=1}^{n} 1 + C_{ik} + \tilde{C}}\right]$$

This function reaches its maximum for the state i that was observed least, as captured by the total number of transitions from that state to any other state, $C_i = \sum_{k=1}^{n} C_{ik}$. Therefore, we can maximize our chances of discovering new states (e.g., transitioning to an as yet undiscovered state) by running simulations from the most poorly sampled states discovered so far. Feature-scaling the number of observations of each state to favor poorly sampled states and to put this undirected component of our reward function on the same scale as the directed component yields

$$\bar{\psi}(i) = \frac{C_{\max} - C_i}{C_{\max} - C_{\min}}$$

where $C_{\min}$ and $C_{\max}$ are the minimum and maximum number of observations of any state, respectively. Favoring poorly sampled states parallels a previously reported heuristic for discovering new conformations.[42] However, we emphasize that balancing this with the directed component of our reward function provides a dramatic improvement in performance, as described in the Results section. The Results section also provides an explicit example of how this works in practice.

To determine how to set the balance between the directed and undirected components of FAST's reward function, the

DOI: 10.1021/acs.jctc.5b00737
J. Chem. Theory Comput. 2015, 11, 5747−5757

algorithm was run with different values of the α parameter using synthetic trajectories generated with existing MSMs, as has been done in previous work on adaptive sampling algorithms.[16] Values ranging from 0.5 to 1.5 gave very similar results, so $\alpha = 1$ was selected to place equal weight on the two components for this study. However, there is no guarantee that this value of α will be optimal for every application. Future work on how best to set this parameter may be valuable.

Simulation parameters for production runs with real molecular dynamics simulations are described below. For β-lactamase, 50 rounds of simulations were run. Each round consisted of a swarm of 30 simulations, each 10 ns in length. Therefore, a total of 15 μs of simulation was run for each variant of FAST performed for this study. For the variant of the villin headpiece, 20 rounds of simulations were run. Each round consisted of a swarm of 10 simulations, each 5 ns in length. Therefore, a total of 1 μs of simulation was run. These simulation lengths were chosen to balance a trade-off between two competing factors: (1) needing simulations to be longer than the lag time used for the final model so that a reasonable MSM can be generated and so that each simulation has a reasonable chance of hopping to a new state and (2) favoring shorter simulations so that each trajectory remains near the region of conformational space where more data is desired rather than drifting to less desirable structures.

MD Simulations. All simulations were run with Gromacs 4.6.5.[43,44] β-Lactamase simulations were run at 300 K using the AMBER ff96 force field[45] with the OBC GBSA implicit solvent model.[46] Using implicit solvent is advantageous for these initial tests as we do not have to store water degrees of freedom or resolvate/re-equilibrate protein conformations when spawning new swarms of simulations. The single starting conformation used for all of these simulations was generated by placing the crystallographic structure of β-lactamase (PDB ID: 1BTL[47]) in a cubic box that extended one nm beyond the protein in any dimension. This system was energy minimized with the steepest descent algorithm until the maximum force fell below 1000 kJ/mol/min using a step size of 0.01 nm and a cutoff distance of 1.2 nm for the neighbor list, Coulomb interactions, and van der Waals interactions. For production runs, all bonds were constrained with the LINCS algorithm[48] and virtual sites[49] were used to allow a 4 fs time step. Cut-offs of 1.0 nm were used for the neighbor list, Coulomb interactions, and van der Waals interactions. The Verlet cutoff scheme was used for the neighbor list. The stochastic velocity rescaling (v-rescale) thermostat[50] was used to hold the temperature at 300 K. Conformations were stored every 10 ps. For the villin headpiece (PDB ID: 2F4K[51]), the simulation settings and one of the extended starting structures from a previous study (structure 5) were employed.[52] Structures were drawn with PyMOL.[53]

Clustering and MSM Construction. All clustering and MSM construction were performed with MSMBuilder.[54,55] An MSM is a discrete-time Master equation model that models protein dynamics as stochastic hopping between discrete conformational states.[39] The states are identified by dividing conformational space up into discrete states, typically by clustering all of the conformations sampled by some set of molecular dynamics simulations. Then, a transition count matrix is constructed, where the element in row i and column j contains the number of transitions from state i to state j observed over the course of some observation interval, called the lag time of the model. The counts matrix is then used to infer a transition probability matrix that contains the probability of transitioning from every possible starting state i to every possible ending state j within a lag time. These matrices are typically estimated with an iterative procedure for identifying the maximum likelihood set of transition probabilities that satisfy microscopic reversibility.[56,57] Thermodynamic and kinetic properties can then be derived from the transition probability matrix rather than the raw simulation data. As a result, these properties are insensitive to the distribution of the starting points used for each simulation, as long as there is sufficient data to obtain a reasonable estimate of the transition probabilities out of each state.[37,38] While building an MSM from the final data set is extremely important for obtaining the proper thermodynamics and kinetics, the clustering of each round of FAST simulations need not be a well-behaved MSM since our reward function does not depend on estimates of the transition probabilities between states. Therefore, these intermediate models just require a clustering with sufficient resolution to detect fluctuations that optimize the target function.

The same clustering procedure was used to analyze each round of simulations and to build an MSM for the final data set. Following a standard protocol,[56] every 10th conformation from the simulations for each protein was clustered with a k-centers algorithm based on the RMSD between protein conformations. The remaining 90% of the data was then assigned to these clusters, and a lag time was selected based on an implied time scales plot.[58] FAST-SASA β-lactamase simulations were clustered based on the RMSD between all backbone heavy atoms and C_β atoms until every cluster had a radius, i.e., maximum distance between any data point in the cluster and the cluster center, less than 1.0 Å and a lag time of 30 ps was employed. FAST-RMSD β-lactamase simulations were clustered based on the RMSD between the helices and loops that move the most when comparing the starting and ending structures (all backbone heavy atoms and C_β atoms in helices 11 and 12 and the loops before and after helix 11, which include residues 215−227 and 270−290) until every cluster had a radius less than 1.0 Å and a lag time of 30 ps was employed. FAST-energy villin simulations were clustered based on the RMSD between all backbone heavy atoms and C_β atoms until every cluster had a radius less than 3.0 Å and a lag time of 2.5 ns was employed. Smaller clusters were employed for the β-lactamase simulations because the conformational changes we intended to capture were subtler than the folding process we targeted in the villin application. The same settings were also used for our retrospective analysis of existing β-lactamase[27] and villin simulations.[52]

Other Analyses. Pocket detection was performed with an implementation of LIGSITE.[27,59] RMSDs and solvent-accessible surface areas were calculated with MDTraj.[60] The highest flux paths between specific starting and ending conformations were performed with transition path theory.[61,62]

RESULTS

Many Physical Properties Have Gradients in Conformational Space. FAST will perform best if the physical property of interest has gradients in conformational space. We hypothesized that the correlation between structures and energies that gives rise to the energetic drive to fold might also give rise to similar gradients in conformational space for other physical properties of proteins. As a first test of this hypothesis, analysis of a number of existing MSMs was

DOI: 10.1021/acs.jctc.5b00737
J. Chem. Theory Comput. 2015, 11, 5747−5757

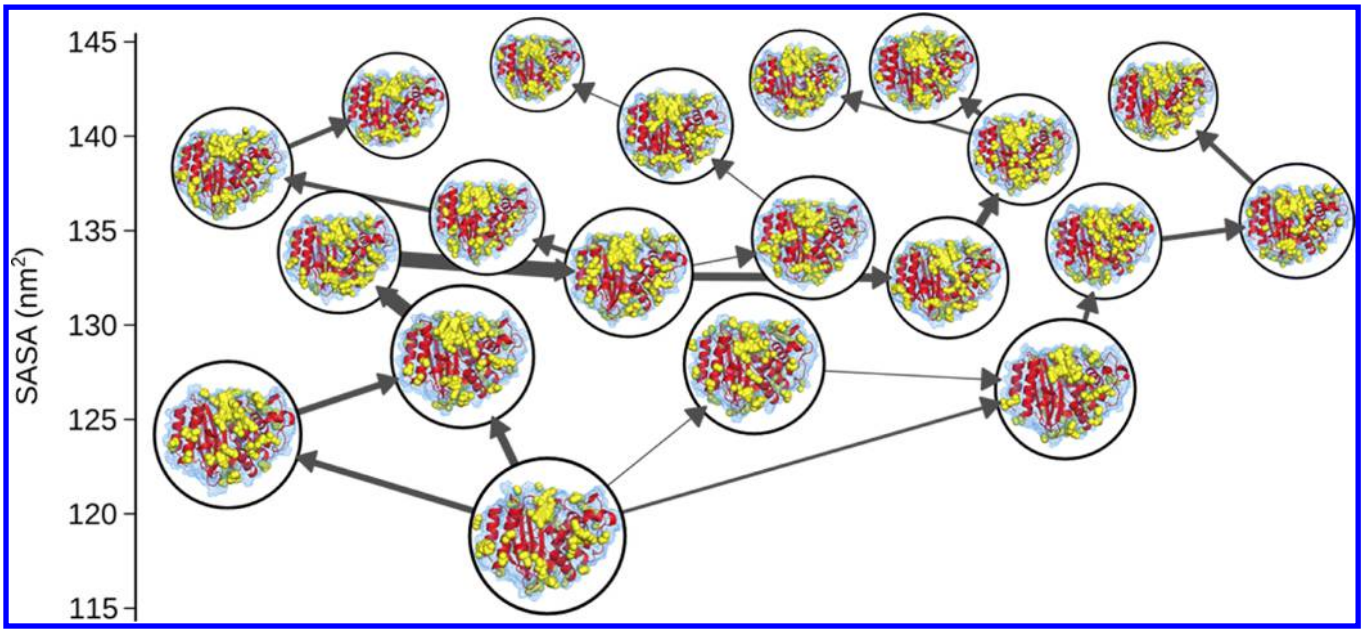


Figure 2. Transition pathways from the crystal structure of TEM-1 β-lactamase to the five states with the largest solvent-accessible surface areas (SASAs) observed in our past work. β-Lactamase is depicted with a red ribbon following the backbone, a blue mesh for the surface, and yellow spheres filling the observed pockets on the protein surface. State sizes are inversely proportional to their free energies, so larger states have higher equilibrium probabilities. Line thickness is directly proportional to the relative flux observed between the start and end states.

performed to determine if the highest flux paths from the crystallographic state to the states that optimize some geometric property do indeed have roughly monotonically increasing (or decreasing) values of that property. For example, Figure 2 shows the preferred pathways from the crystal structure of TEM-1 β-lactamase to the states with the highest solvent-accessible surface areas discovered in 81 μs of aggregate simulation conducted on the Folding@home distributed computing environment.[27] The solvent-accessible surface areas of structural states along these high-flux pathways tend to increase monotonically, so it is reasonable to expect the directed component of FAST to help the algorithm move along these paths quickly. There are some backward steps along these paths that require moving from states with larger solvent-accessible surface areas to states with lower surface areas, but these steps are small enough that it is also reasonable to expect the undirected, statistical component of FAST to easily overcome these hurdles. Similar trends are also observed for properties like the energy and RMSD to a selected target structure in this model of β-lactamase as well as models for proteins like a fast-folding variant of the villin headpiece (500 μs of simulation),[52] NTL9 (1.5 ms of simulation),[25] and lambda repressor (1.3 ms of simulation).[63] Taken together, this evidence supports the hypothesis that many physical properties have gradients in conformational space that the FAST algorithm is intended to identify and follow.

FAST Accurately Identifies the Preferred Paths to Target Conformations. If FAST works as intended, then it should be capable of quickly following gradients in conformational space to find the preferred paths to structures that optimize a selected geometric function. As a first test of whether FAST successfully achieves this goal, we compared its performance to conventional simulations using an existing MSM to generate synthetic trajectories via kinetic Monte Carlo. To generate a synthetic trajectory, one first selects a starting state, then uses the transition probabilities out of that state to randomly select a new state, and repeats this procedure until a desired trajectory length is reached. Synthetic trajectories can then be used to estimate the transition probabilities between states to reconstruct the MSM with which they were generated. Performing initial tests with such synthetic trajectories is advantageous because (1) it is much more computationally efficient than running real molecular dynamics simulations and (2) the MSM used to generate the trajectories serves as a gold standard for assessing the performance of different methods.

We chose a previously reported relative entropy metric to assess the quality of MSMs reconstructed from synthetic trajectories.[16] The relative entropy between two MSMs is

$$D(P \| Q) = \sum_{i,j}^{N} P_i P_{ij} \log \frac{P_{ij}}{Q_{ij}}$$

where P is the transition matrix for the reference MSM used to generate the synthetic trajectories, P_i is the equilibrium probability of state i in that MSM, and P_{ij} is the probability of hopping from state i to state j in the reference MSM. Q, Q_i, and Q_{ij} are the corresponding properties of the MSM reconstructed from synthetic trajectories. The relative entropy is zero if the two MSMs are identical and becomes increasingly large the more the two models differ. To ensure that every transition probability is nonzero and avoid infinite relative entropies, we used a pseudocount of $1/n$, where n is the number of states in the model, as described in the Methods section and our previous work.[16]

We used our existing MSM for β-lactamase to simulate how quickly FAST-RMSD finds structures resembling conformations bound to a surprising allosteric ligand compared to that using conventional simulations. First, we identified the five states with the lowest RMSD to the target structure and identified the three highest flux pathways from the state containing the ligand-free crystal structure to each of the five target states (15 paths total). Together, these paths contained 32 of the 3469 states in the MSM. Then, we ran long

DOI: 10.1021/acs.jctc.5b00737
J. Chem. Theory Comput. 2015, 11, 5747−5757

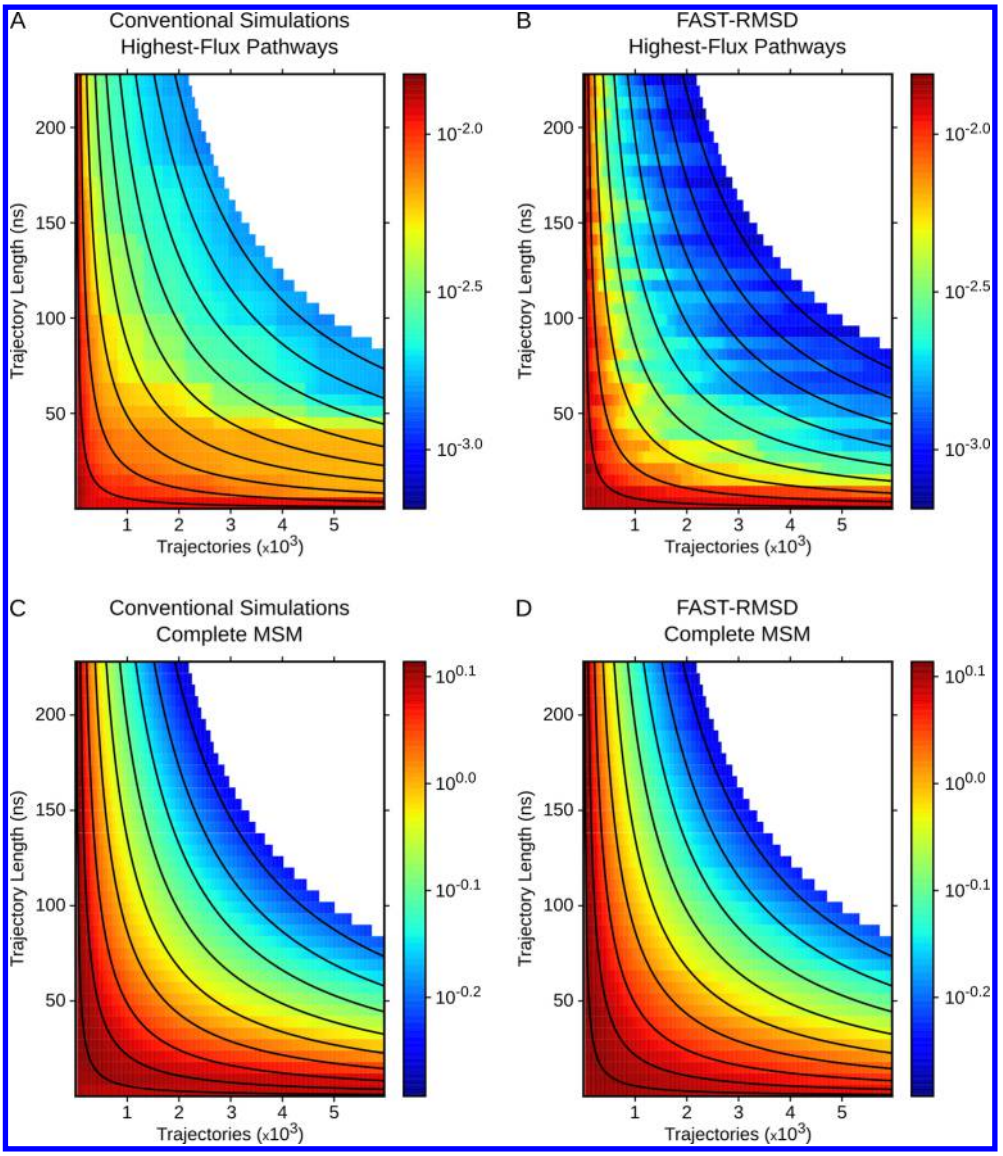


Figure 3. Relative entropies between the gold-standard MSM of β-lactamase and MSMs constructed with different sampling methods using varying numbers of kinetic Monte Carlo simulations of different lengths. Panels A and B show the relative entropies for a subset of states along the highest flux pathways to the five states with the lowest RMSDs to a target structure for conventional and FAST-RMSD simulations, respectively. Panels C and D show the relative entropies for the entire MSMs from each sampling method. Black contours indicate equivalent aggregate simulation time. Calculations were not performed for the white regions.

conventional simulations and FAST-RMSD simulations started from the crystallographic state, constructed MSMs from each set of synthetic trajectories, and employed the relative entropy metric to assess how well each method captured the transition probabilities for the 32 states along the highest-flux pathways to low RMSD states. Figure 3A,B shows the results of repeating this analysis for varying numbers of simulations of different lengths. These results demonstrate that FAST-RMSD accurately captures this structural subspace with far less total simulation time than conventional simulations. Comparing the methods across all states (Figure 3C,D) also demonstrates that FAST yields models that are as accurate as conventional simulations on a global level.

Together, these results suggest that it is possible to extract the proper thermodynamics and kinetics from FAST simulations despite the fact that starting points for simulations are not chosen according to a Boltzmann distribution. As a further test of the algorithm, we also applied it to three real-world problems using real molecular dynamics simulations instead of synthetic trajectories, as described below.

FAST-SASA Discovers a Diversity of Pocket Structures. One use of molecular dynamics simulations is to discover unexpected pockets that open as a protein fluctuates away from its crystal structure that might serve as valuable drug targets. Since the opening of pockets will generally increase a protein structure's solvent-accessible surface area,[64] we chose to maximize this property using FAST-SASA.

To understand how FAST works, the highest-flux pathways from the initial (crystallographic) state to the five states with the largest solvent-accessible surface areas discovered by FAST-SASA were identified and colored according to when they were first discovered, as shown in Figure 4. In the first few rounds of simulation, FAST-SASA finds a few states with somewhat higher solvent-accessible surface areas, such as states A and B. At this point, these states have the highest solvent-accessible surface areas and are poorly sampled since they were just

DOI: 10.1021/acs.jctc.5b00737
J. Chem. Theory Comput. 2015, 11, 5747−5757

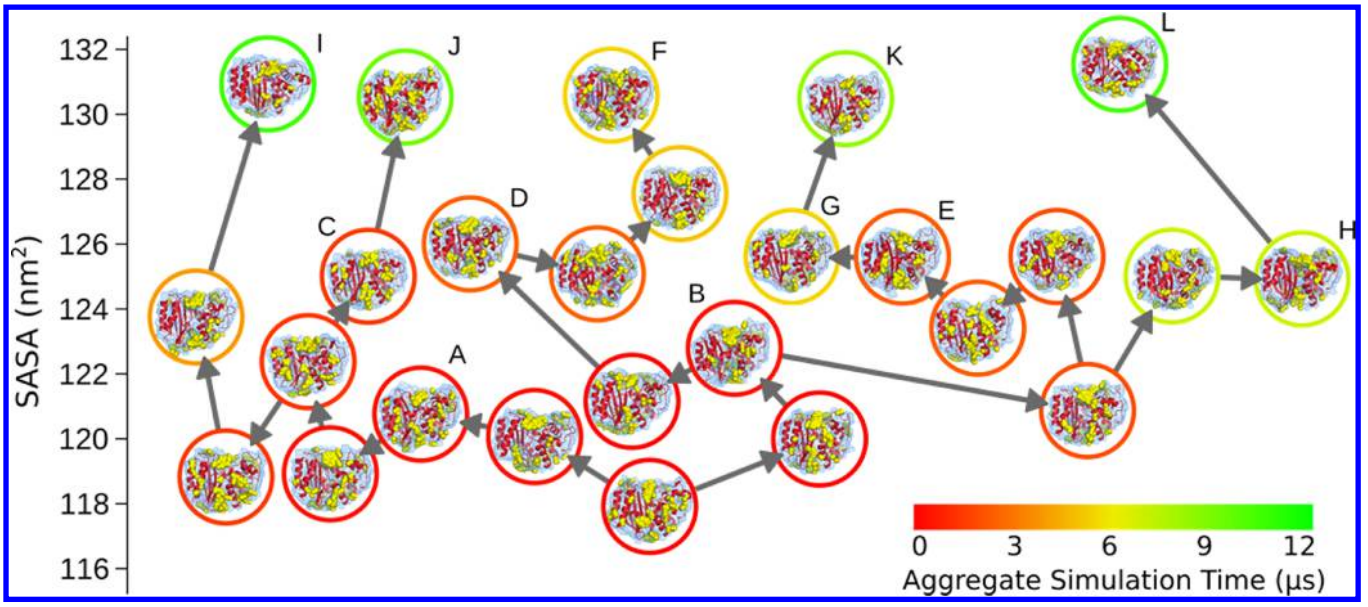


Figure 4. Transition pathways from the crystal structure of TEM-1 β-lactamase to the five states with the largest solvent-accessible surface areas (SASAs) discovered using FAST-SASA. β-Lactamase is depicted with a red ribbon following the backbone, a blue mesh for the surface, and yellow spheres filling the observed pockets on the protein surface. States are colored to indicate when they were discovered during the course of 12 μs of FAST-SASA sampling.

discovered. Therefore, they are selected as starting conformations for the next round of simulations. Simulations spawned from these states then discover states C–E, which are selected as the starting points for the next swarm, again because they have large solvent-accessible surface areas and are poorly sampled. Simulations that are spawned from state D, and those subsequently discovered, lead to the discovery of state F, one of the states with the largest solvent-accessible surface areas. When the sampling of state F fails to produce new states with larger solvent-accessible surface areas, its ranking decreases, leading to the favoring of states that have been sampled less despite having a lower solvent-accessible surface area. Sampling from these lower-solvent-accessible surface areas helps to discover a variety of new states, such as states G and H, that have the potential to elucidate new pathways to high solvent-accessible surface area states. These states are ranked highly due to their recent discovery and manage to discover independent pathways to some of the other states with the largest solvent-accessible surface areas (I–L). The yellow spheres in Figure 4 fill in pockets that open in the protein structures, highlighting that there are distinct pockets forming in different states with equivalent solvent-accessible surface areas.

To assess the performance of FAST-SASA, we compared it to conventional molecular dynamics simulations, a purely SASA-based sampling scheme that uses just the directed component of FAST-SASA, and a variant of counts-based adaptive sampling that uses just the undirected component of FAST-SASA. An equivalent amount of conventional molecular dynamics simulations (ten 1.5 μs simulations) explore only conformations near the crystal structure, as shown in Figure 5A. The small increases in solvent-accessible surface area that these simulations achieve make a quantitative comparison with FAST-SASA impossible, so we can conclude only that FAST-SASA is orders of magnitude more efficient.

Counts-based sampling is also significantly less efficient than FAST-SASA. The fact that this algorithm lacks a directed component prevents it from aggressively capitalizing on

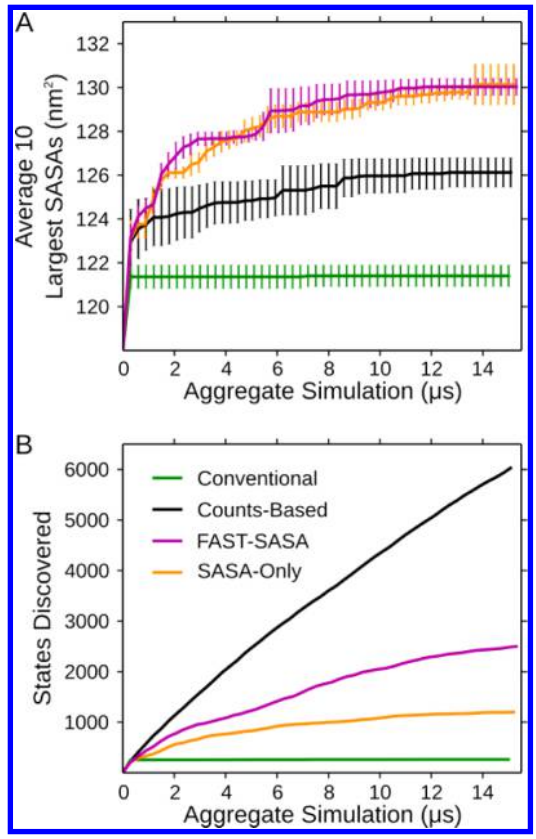


Figure 5. Performance of FAST-SASA (magenta) compared to conventional molecular dynamics (green), count-based sampling (black), and SASA-based sampling (orange). (A) Average of the solvent-accessible surface areas for the 10 states with the largest surface areas discovered as a function of aggregate simulation time. (B) Number of states discovered as a function of the aggregate simulation time.

DOI: 10.1021/acs.jctc.5b00737
J. Chem. Theory Comput. 2015, 11, 5747–5757

promising structures. Instead, counts-based sampling tries to build out from every new state that it discovers. In doing so, it discovers more total states than FAST-SASA, as shown in Figure 5B, but most have small solvent-accessible surface areas and do not have the sort of pockets that we set out to discover in this application. FAST-SASA finds states with equally large solvent-accessible surface areas at least 8 times faster than counts-based sampling alone. Moreover, this is a conservative estimate of the improved performance of FAST-SASA because it finds at least 30 times as many conformations with surface areas greater than 125 nm^2. Finding equivalent diversity with counts-based sampling would likely take orders of magnitude more simulation than with FAST-SASA given the undirected nature of the purely counts-based algorithm.

SASA-based sampling finds states with much higher solvent-accessible surface areas than the conventional simulations or counts-based sampling (Figure 5A). Indeed, SASA-based simulations find a few states with solvent-accessible surface areas that are comparable to the best structures found by FAST-SASA. However, compared to FAST-SASA, it essentially finds a single a high solvent-accessible surface area state and then persistently simulates that state because it lacks the undirected component that allows FAST-SASA to give up on a state and reroute to other potentially more fruitful starting conformations. Therefore, FAST-SASA discovers far more states (Figure 5B), including at least twice as many conformations with surface areas greater than 125 nm^2. Since SASA-based sampling persistently spawns new simulations from the single high surface area state that it finds, it is unlikely to ever discover the diversity of structures that FAST-SASA finds. Therefore, as with the conventional simulations, we conclude that FAST-SASA is orders of magnitude more efficient.

FAST-RMSD Efficiently Finds Paths between Specific Structures. Computer simulations are also frequently employed to discover the transition paths between two distinct structures. As an example of this sort of problem, we sought to discover the preferred paths from the ligand-free crystal structure of β-lactamase discussed in the previous section to a structure with an unexpected allosteric binding pocket (1PZO[65]). To accelerate the discovery of such paths, we used FAST-RMSD to discover structures with low RMSDs to the target structure and compared the performance of these simulations to conventional molecular dynamics simulations and counts-based adaptive sampling. All of the trends are similar to those observed for FAST-SASA in comparison to other sampling methods, as shown in Figure 6. Combined with our analysis of synthetic trajectories, as described earlier, we conclude that FAST-RMSD quickly finds target structures and the preferred paths to these structures.

FAST-Energy Folds Proteins. As a final test of FAST, we applied it to the folding of a variant of the villin headpiece that folds in ~700 ns.[51] Inspired by the idea that proteins fold by following an energy gradient toward their native states, we chose to run FAST-energy to minimize the system's energy. This choice also allows *bona fide* structure predictions rather than building in the answer with a method like FAST-RMSD. To make a direct comparison with a past study of this protein conducted on the Folding@home distributed computing environment,[52] the same simulation parameters and explicit solvent were used. However, the energies used in FAST's reward function were calculated using implicit solvent because water−water interactions will dominate the energy of any

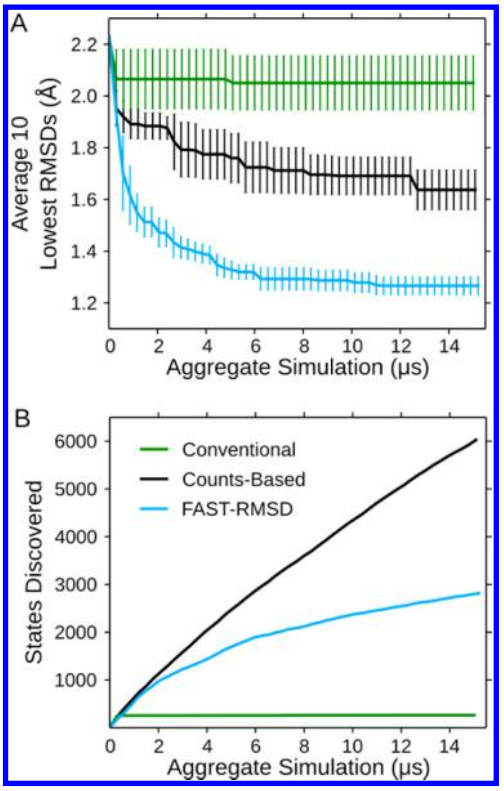


Figure 6. Performance of FAST-RMSD (cyan) compared to conventional molecular dynamics (green) and count-based sampling (black). (A) Average of the RMSD to the target structure for the 10 states with the lowest RMSDs discovered as a function of aggregate simulation time. (B) Number of states discovered as a function of the aggregate simulation time.

structure with explicit solvent. Implicit solvent, on the other hand, integrates out the water degrees of freedom, allowing FAST-energy to focus on finding preferred protein structures.

FAST-energy simulations fold villin to within 2.5 Å of its crystal structure in just 400 ns of aggregate simulation. Figure 7 state A shows the extended starting structure used for these simulations, and Figure 7 state B shows the predicted structure overlaid on the crystal structure. This result is impressive because there is only a ~60% chance of folding the protein with 700 ns of conventional simulation based on the experimental folding time. Furthermore, the previous Folding@home study that inspired our FAST-energy calculations used 500 μs of conventional simulation,[52] and a folding study run on the ANTON supercomputer used 125 μs of simulation.[26]

To understand the structural ensemble explored by FAST-energy, scatter plots of the energies of states from the MSM built from the FAST-energy data vs their RMSDs to the crystal structure were overlaid with the same information from past Folding@home studies,[52,56] as shown in Figure 7. Overall, the model from FAST-energy covers a similar range of energies and RMSDs as that found by conventional molecular dynamics simulations. However, visual inspection of the scatter plot suggests that FAST-energy finds more structures with both low energies and low RMSDs. This observation is further supported by the histograms of the energies and RMSDs for the structural states discovered by each method. Taken together, these results demonstrate that FAST successfully discovers the energetically accessible conformations that would eventually be found by

DOI: 10.1021/acs.jctc.5b00737
J. Chem. Theory Comput. 2015, 11, 5747−5757

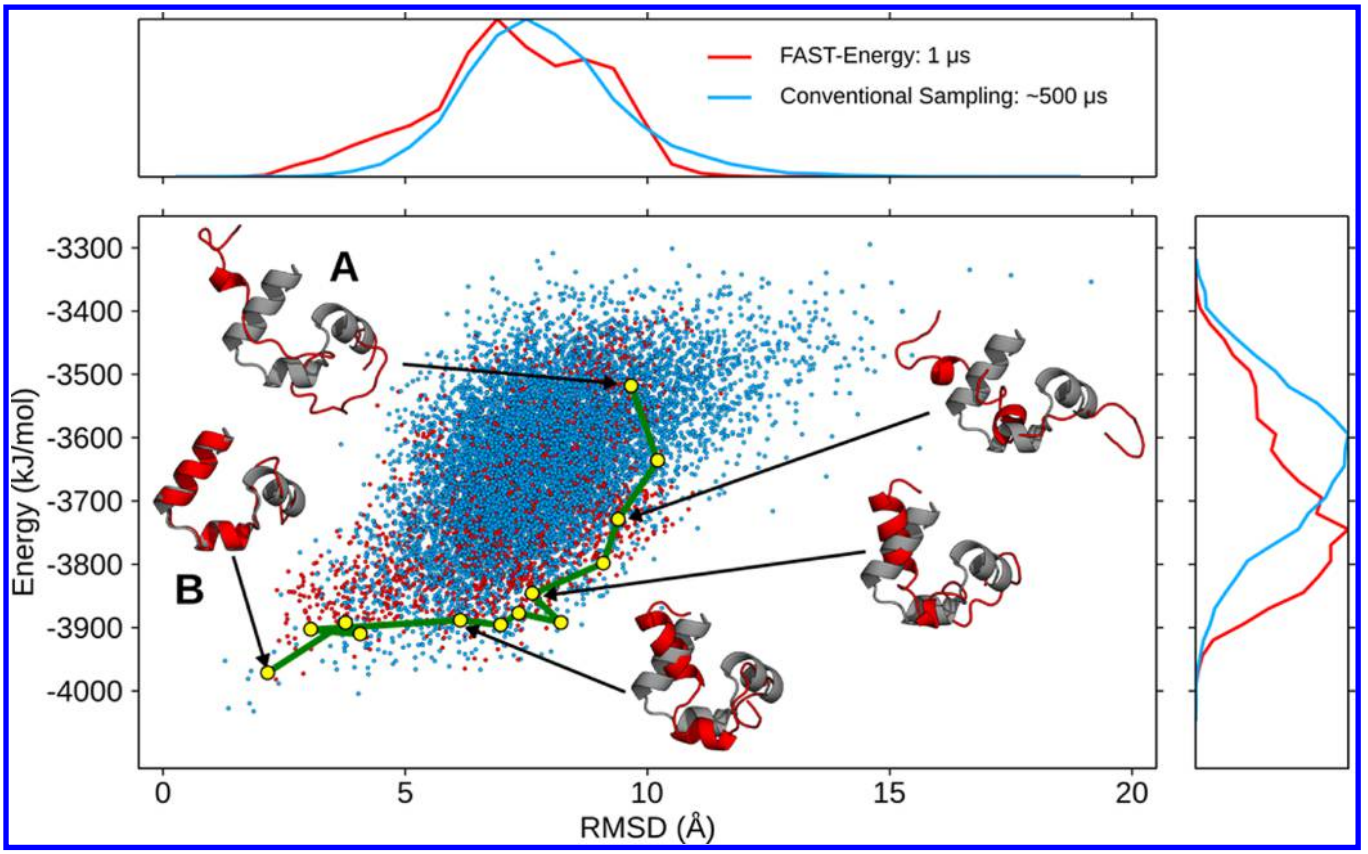


Figure 7. State−space of villin projected onto two order parameters: potential energy and the RMSD to the native crystal structure. Each point represents a single state discovered within 1 μs of FAST-energy sampling (red) or 500 μs of unguided sampling from Folding@home (blue). Normalized histograms of the number of states with a given potential energy (right plot) or RMSD (top plot) are shown. The highest-flux pathway from the unfolded starting state (state A) to the state with the lowest RMSD (state B) is plotted as a green line, where states along the pathway are identified with yellow points. Five conformations along the FAST-energy folding pathway (red) are superimposed onto the native crystal structure (gray).

conventional simulations, but it does so with much less simulation.

To see if FAST-energy finds similar folding routes to past studies or if the reward function used to choose starting structures for each round of simulation somehow biases the result, the preferred folding pathway from the final MSM was identified. This model ought to capture the proper thermodynamics and kinetics of the states visited.[37,38] Indeed, the protein first forms some elements of secondary structure and begins to collapse, as observed in previous studies.[66,67] The N-terminal helix is also the last to form, in agreement with previous studies using the same force field.[68] Finally, the slowest implied time scale of the model was calculated as an estimate of the folding time. This calculation yielded a folding time of 830 ± 260 ns, again in reasonable agreement with both experiment and past work using conventional molecular dynamics simulations. Therefore, we conclude that MSMs built from FAST simulations are indeed capable of capturing the proper thermodynamics and kinetics despite the fact that starting conformations are not selected according to a Boltzmann distribution.

■ CONCLUSIONS

We have introduced a goal-oriented sampling method, called FAST, which rapidly searches through conformational space for structures with desired properties by balancing exploration/exploitation trade-offs. This algorithm was inspired by the hypothesis that many physical properties have an overall gradient in conformational space, akin to the energetic gradients that are known to guide proteins to their folded states. Indeed, retrospective analysis of existing MSMs supports the idea that structural properties like the RMSD to a target structure, the solvent-accessible surface area, and the energy have such gradients. To follow these gradients, we designed FAST to balance between (1) recognizing and amplifying small motions that maximize (or minimize) a selected geometric function and (2) exploring poorly sampled regions of configuration space. This balance is achieved by leveraging ideas from optimization theory regarding exploration/exploitation trade-offs.

To test FAST, we applied it to a number of common problems and compared its performance to alternative approaches, such as conventional molecular dynamics simulations and counts-based adaptive sampling. For example, we demonstrated that FAST can find pockets by preferentially sampling structures with large surface areas, it can find paths between specific structures by minimizing the RMSD to a target, and it can fold proteins by minimizing their energies. In each case, FAST outperforms the methods to which we compared it by at least an order of magnitude and likely considerably more. The success of FAST supports our hypothesis that many physical properties have gradients in conformational space. Moreover, our results demonstrate that FAST is capable of identifying and following these gradients,

DOI: 10.1021/acs.jctc.5b00737
J. Chem. Theory Comput. 2015, 11, 5747−5757

even overcoming and circumventing barriers that interrupt these trends. In addition to finding structures with a desired property more quickly than other algorithms, FAST also finds a greater diversity of such structures. While the data generated with FAST is not Boltzmann distributed, building an MSM from the data provides the proper thermodynamics and kinetics. The ability to obtain broad sampling while maintaining the proper kinetics is an important advantage over many other sampling algorithms that facilitates a direct connection with kinetic experiments. Therefore, we expect FAST to be of great utility for a wide range of applications. There are also many opportunities for combining FAST with other sampling methods. For example, one could use accelerated molecular dynamics to obtain even broader sampling, though this would sacrifice kinetics. One could also use FAST for state discovery and then refine estimates of the transition probabilities between states with adaptive sampling schemes designed to reduce statistical uncertainty.

AUTHOR INFORMATION

Corresponding Author

*E-mail: bowman@biochem.wustl.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We are grateful to Chris Ho for helpful suggestions to the figures. G.R.B. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund.

REFERENCES

(1) Israelewitz, B.; Gao, M.; Schulten, K. Steered Molecular Dynamics and Mechanical Functions of Proteins. *Curr. Opin. Struct. Biol.* **2001**, *11* (2), 224−230.

(2) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (20), 12562−12566.

(3) Huber, T.; Torda, A. E.; van Gunsteren, W. F. Local Elevation: a Method for Improving the Searching Properties of Molecular Dynamics Simulation. *J. Comput.-Aided Mol. Des.* **1994**, *8* (6), 695−708.

(4) Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. String Method in Collective Variables: Minimum Free Energy Paths and Isocommittor Surfaces. *J. Chem. Phys.* **2006**, *125* (2), 024106.

(5) Pan, A. C.; Sezer, D.; Roux, B. Finding Transition Pathways Using the String Method with Swarms of Trajectories. *J. Phys. Chem. B* **2008**, *112* (11), 3432−3440.

(6) MacCallum, J. L.; Perez, A.; Dill, K. A. Determining Protein Structures by Combining Semireliable Data with Atomistic Physical Models by Bayesian Inference. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 6985−6990.

(7) Lindorff-Larsen, K.; Best, R. B.; DePristo, M. A.; Dobson, C. M.; Vendruscolo, M. Simultaneous Determination of Protein Structure and Dynamics. *Nature* **2005**, *433* (7022), 128−132.

(8) Zheng, L.; Chen, M.; Yang, W. Random Walk in Orthogonal Space to Achieve Efficient Free-Energy Simulation of Complex Systems. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (51), 20227−20232.

(9) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314* (1−2), 141−151.

(10) Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated Molecular Dynamics: a Promising and Efficient Simulation Method for Biomolecules. *J. Chem. Phys.* **2004**, *120* (24), 11919.

(11) Huber, G. A.; Kim, S. Weighted-Ensemble Brownian Dynamics Simulations for Protein Association Reactions. *Biophys. J.* **1996**, *70* (1), 97−110.

(12) Dickson, A.; Brooks, C. L. WExplore: Hierarchical Exploration of High-Dimensional Spaces Using the Weighted Ensemble Algorithm. *J. Phys. Chem. B* **2014**, *118* (13), 3532−3542.

(13) Suárez, E.; Lettieri, S.; Zwier, M. C.; Stringer, C. A.; Subramanian, S. R.; Chong, L. T.; Zuckerman, D. M. Simultaneous Computation of Dynamical and Equilibrium Information Using a Weighted Ensemble of Trajectories. *J. Chem. Theory Comput.* **2014**, *10* (7), 2658−2667.

(14) Chen, Y.; Roux, B. Efficient Hybrid Non-Equilibrium Molecular Dynamics–Monte Carlo Simulations with Symmetric Momentum Reversal. *J. Chem. Phys.* **2014**, *141* (11), 114107.

(15) Hinrichs, N.; Pande, V. Calculation of the Distribution of Eigenvalues and Eigenvectors in Markovian State Models for Molecular Dynamics. *J. Chem. Phys.* **2007**, *126*, 244101.

(16) Bowman, G. R.; Ensign, D.; Pande, V. Enhanced Modeling via Network Theory: Adaptive Sampling of Markov State Models. *J. Chem. Theory Comput.* **2010**, *6*, 787−794.

(17) Doerr, S.; De Fabritiis, G. On-the-Fly Learning and Sampling of Ligand Binding by High-Throughput Molecular Simulations. *J. Chem. Theory Comput.* **2014**, *10* (5), 2064−2069.

(18) Bacci, M.; Vitalis, A.; Caflisch, A. A Molecular Simulation Protocol to Avoid Sampling Redundancy and Discover New States. *Biochim. Biophys. Acta, Gen. Subj.* **2015**, *1850* (5), 889−902.

(19) Adhikari, A. N.; Freed, K. F.; Sosnick, T. R. Simplified Protein Models: Predicting Folding Pathways and Structure Using Amino Acid Sequences. *Phys. Rev. Lett.* **2013**, *111* (2), 028103.

(20) Voelz, V. A.; Elman, B.; Razavi, A. M.; Zhou, G. Surprisal Metrics for Quantifying Perturbed Conformational Dynamics in Markov State Models. *J. Chem. Theory Comput.* **2014**, *10* (12), 5716−5728.

(21) Moyano, G. E.; Collins, M. A. Molecular Potential Energy Surfaces by Interpolation: Strategies for Faster Convergence. *J. Chem. Phys.* **2004**, *121* (20), 9769−9775.

(22) Shirts, M.; Pande, V. S. COMPUTING: Screen Savers of the World Unite! *Science* **2000**, *290* (5498), 1903−1904.

(23) Buch, I.; Harvey, M. J.; Giorgino, T.; Anderson, D. P.; De Fabritiis, G. High-Throughput All-Atom Molecular Dynamics Simulations Using Distributed Computing. *J. Chem. Inf. Model.* **2010**, *50* (3), 397−403.

(24) Shaw, D.; Dror, R.; Salmon, J.; Grossman, J. P.; Mackenzie, K. M.; Bank, J. A.; Young, C.; Deneroff, M. M.; Batson, B.; Bowers, K. J.; Eastwood, M. P.; Ierardi, D. J.; Klepeis, J. L.; Kuskin, J. S.; Larson, R. H.; Lindorff-Larsen, K.; Maragakis, P.; Moraes, M. A.; Piana, S.; Shan, Y.; Towles, B. Millisecond-Scale Molecular Dynamics Simulations on Anton. Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, Portland, OR, November 14−20, 2009.

(25) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. Molecular Simulation of Ab Initio Protein Folding for a Millisecond Folder NTL9 (1− 39). *J. Am. Chem. Soc.* **2010**, *132* (5), 1526−1528.

(26) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **2011**, *334* (6055), 517−520.

(27) Bowman, G. R.; Geissler, P. L. Equilibrium Fluctuations of a Single Folded Protein Reveal a Multitude of Potential Cryptic Allosteric Sites. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (29), 11681−11686.

(28) Dror, R. O.; Green, H. F.; Valant, C.; Borhani, D. W.; Valcourt, J. R.; Pan, A. C.; Arlow, D. H.; Canals, M.; Lane, J. R.; Rahmani, R.; Baell, J. B.; Sexton, P. M.; Christopoulos, A.; Shaw, D. E. Structural Basis for Modulation of a G-Protein-Coupled Receptor by Allosteric Drugs. *Nature* **2013**, *50*, 295−299.

(29) Kohlhoff, K. J.; Shukla, D.; Lawrenz, M.; Bowman, G. R.; Konerding, D. E.; Belov, D.; Altman, R. B.; Pande, V. S. Cloud-Based Simulations on Google Exacycle Reveal Ligand Modulation of GPCR Activation Pathways. *Nat. Chem.* **2013**, *6* (1), 15−21.

(30) Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. Theory of Protein Folding: the Energy Landscape Perspective. *Annu. Rev. Phys. Chem.* **1997**, *48* (1), 545−600.

DOI: 10.1021/acs.jctc.5b00737
J. Chem. Theory Comput. 2015, 11, 5747−5757

(31) Brooks, C. L. Simulations of Protein Folding and Unfolding. *Curr. Opin. Struct. Biol.* **1998**, *8* (2), 222−226.
(32) Dill, K. A.; Chan, H. S. From Levinthal to Pathways to Funnels. *Nat. Struct. Biol.* **1997**, *4* (1), 10−19.
(33) Berry, D. A.; Fristedt, B. *Bandit Problems*; Chapman and Hall: London, 1985.
(34) Poli, R.; Kennedy, J.; Blackwell, T. Particle Swarm Optimization. *Swarm Intell* **2007**, *1* (1), 33−57.
(35) Audibert, J.-Y.; Munos, R.; Szepesvári, C. Exploration−Exploitation Tradeoff Using Variance Estimates in Multi-Armed Bandits. *Theor Comput. Sci.* **2009**, *410* (19), 1876−1902.
(36) Auer, P. Using Confidence Bounds for Exploitation-Exploration Trade-Offs. *J. Mach. Learn. Res.* **2003**, *3*, 397−422.
(37) Huang, X.; Bowman, G. R.; Bacallado, S.; Pande, V. S. Rapid Equilibrium Sampling Initiated From Nonequilibrium Data. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (47), 19765−19769.
(38) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the Equilibrium Ensemble of Folding Pathways From Short Off-Equilibrium Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (45), 19011−19016.
(39) *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*; Bowman, G. R., Pande, V. S., Noé, F., Eds.; Springer: Dordrecht, The Netherlands, 2014.
(40) Noé, F. Probability Distributions of Molecular Observables Computed From Markov Models. *J. Chem. Phys.* **2008**, *128* (24), 244103.
(41) Bowman, G. R. Improved Coarse-Graining of Markov State Models via Explicit Consideration of Statistical Uncertainty. *J. Chem. Phys.* **2012**, *137* (13), 134111.
(42) Weber, J. K.; Pande, V. S. Characterization and Rapid Sampling of Protein Folding Markov State Model Topologies. *J. Chem. Theory Comput.* **2011**, *7* (10), 3405−3411.
(43) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* **2005**, *26* (16), 1701−1718.
(44) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: a High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinformatics* **2013**, *29* (7), 845−854.
(45) Kollman, P. A. Advances and Continuing Challenges in Achieving Realistic and Predictive Simulations of the Properties of Organic and Biological Molecules. *Acc. Chem. Res.* **1996**, *29*, 461−469.
(46) Onufriev, A.; Bashford, D.; Case, D. A. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins: Struct., Funct., Genet.* **2004**, *55* (2), 383−394.
(47) Jelsch, C.; Mourey, L.; Masson, J. M.; Samama, J. P. Crystal Structure of Escherichia Coli TEM1 Beta-Lactamase at 1.8 a Resolution. *Proteins: Struct., Funct., Genet.* **1993**, *16* (4), 364−383.
(48) Hess, B. P-LINCS: a Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4* (1), 116−122.
(49) Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. Improving Efficiency of Large Time-Scale Molecular Dynamics Simulations of Hydrogen-Rich Systems. *J. Comput. Chem.* **1999**, *20* (8), 786−798.
(50) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling Through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126* (1), 014101.
(51) Kubelka, J.; Chiu, T. K.; Davies, D. R.; Eaton, W. A.; Hofrichter, J. Sub-Microsecond Protein Folding. *J. Mol. Biol.* **2006**, *359* (3), 546−553.
(52) Ensign, D.; Kasson, P.; Pande, V. Heterogeneity Even at the Speed Limit of Folding: Large-Scale Molecular Dynamics Study of a Fast-Folding Variant of the Villin Headpiece. *J. Mol. Biol.* **2007**, *374*, 806−816.
(53) DeLano, W. L. *PyMOL Molecular Graphics System*; DeLano Scientific: Palo Alto, CA, 2002
(54) Bowman, G. R.; Huang, X.; Pande, V. S. Using Generalized Ensemble Simulations and Markov State Models to Identify Conformational States. *Methods* **2009**, *49* (2), 197−201.
(55) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. MSMBuilder2: Modeling Conformational Dynamics at the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.* **2011**, *7* (10), 3412−3419.
(56) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. Progress and Challenges in the Automated Construction of Markov State Models for Full Protein Systems. *J. Chem. Phys.* **2009**, *131* (12), 124101.
(57) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov Models of Molecular Kinetics: Generation and Validation. *J. Chem. Phys.* **2011**, *134* (17), 174105.
(58) Swope, W. C.; Pitera, J. W.; Suits, F. Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory †. *J. Phys. Chem. B* **2004**, *108* (21), 6571−6581.
(59) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: Automatic and Efficient Detection of Potential Small Molecule-Binding Sites in Proteins. *J. Mol. Graphics Modell.* **1997**, *15* (6), 359−363.
(60) McGibbon, R. T.; Beauchamp, K. A.; Schwantes, C. R.; Wang, L.-P.; Hernández, C. X.; Harrigan, M. P.; Lane, T. J.; Swails, J. M.; Pande, V. S. MDTraj: a Modern, Open Library for the Analysis of Molecular Dynamics Trajectories. *bioRxiv* **2014**, 008896.
(61) Weinan, E.; Vanden-Eijnden, E. Toward a Theory of Transition Paths. *J. Stat. Phys.* **2006**, *123*, 503−523.
(62) Metzner, P.; Schuette, C.; Vanden-Eijnden, E. Transition Path Theory for Markov Jump Processes. *Multiscale Model. Simul.* **2009**, *7*, 1192−1219.
(63) Bowman, G. R.; Voelz, V. A.; Pande, V. S. Atomistic Folding Simulations of the Five-Helix Bundle Protein Λ(6−85). *J. Am. Chem. Soc.* **2011**, *133* (4), 664−667.
(64) Bowman, G. R.; Bolin, E. R.; Hart, K. M.; Maguire, B. C.; Marqusee, S. Discovery of Multiple Hidden Allosteric Sites by Combining Markov State Models and Experiments. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (9), 2734−2739.
(65) Horn, J. R.; Shoichet, B. K. Allosteric Inhibition Through Core Disruption. *J. Mol. Biol.* **2004**, *336*, 1283−1291.
(66) Duan, Y.; Wang, L.; Kollman, P. A. The Early Stage of Folding of Villin Headpiece Subdomain Observed in a 200-ns Fully Solvated Molecular Dynamics Simulation. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95* (17), 9897−9902.
(67) Bowman, G. R.; Pande, V. S. Protein Folded States Are Kinetic Hubs. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (24), 10890−10895.
(68) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophys. J.* **2011**, *100* (9), L47−L49.

DOI: 10.1021/acs.jctc.5b00737
J. Chem. Theory Comput. 2015, 11, 5747−5757