JCTC Journal of Chemical Theory and Computation

Article

Cite This: J. Chem. Theory Comput. 2018, 14, 2109−2116

pubs.acs.org/JCTC

MELD-Path Efficiently Computes Conformational Transitions, Including Multiple and Diverse Paths

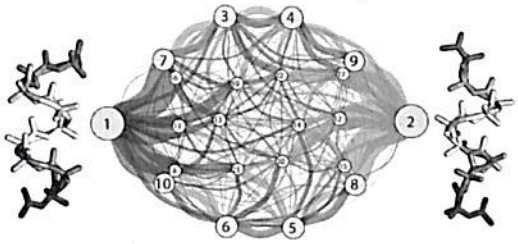
Alberto Perez,[†,§] Florian Sittel,[‡,§] Gerhard Stock,[‡] and Ken Dill*[,†]

[†]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York 11794, United States

[‡]Biomolecular Dynamics, Institute of Physics, Albert Ludwigs University, 79104 Freiburg, Germany

Supporting Information

ABSTRACT: The molecular actions of proteins occur along reaction coordinates. Current computer methods have limited ability to explore them. We describe a fast protocol called MELD-path that (1) efficiently samples relevant conformational states via MELD, an accelerator of Molecular Dynamics (MD), (2) seeds multiple short MD trajectories from MELD states, and then (3) constructs Markov State Models (MSM) that give the routes and kinetics. We tested the method against extensive (multi μs) MD simulations of the right-handed- to left-handed-helix transition of a 9-mer peptide of AIB, the symmetry of which allows us to establish convergence. MELD-path finds all the metastable states, their correct relative populations, and the full ensemble of routes, not just a single assumed route. For this transition, we find a very broad route structure. MELD-path is highly parallelizable and efficient, yielding the full route map in a few days of computation. We believe MELD-path could be a general and rapid way to explore mechanistic processes in biomolecules on the computer.



1. INTRODUCTION

A main way to study the detailed actions and mechanisms of biomolecules is by Molecular Dynamics (MD) computer simulations. Based on the underlying physical driving forces, they can give the picosecond-by-picosecond and Ångstrom-by-Ångstrom narratives that experiments are too coarse-grained to provide. However, MD modeling of biomolecular mechanisms is currently limited, for the following reasons. MD sampling is very inefficient by itself to sample large conformational changes and overcoming kinetic barriers. MD mechanistic modeling requires knowing a proper reaction coordinate, which is often difficult to determine. It must sample the conformations well enough along the reaction coordinates to get accurate closely spaced free-energy distributions, but such computations are prohibitively expensive for all but the simplest problems. Consequently, it is unable to explore more than one or a few dominant routes, even though biomolecule transition routes are likely to be many and varied. We describe here a method called MELD-path that can address these problems, and we give a proof-of-principle example.

It is often of interest to learn about a particular transition between two states (A and B). Sometimes it is possible to guess or identify a reaction coordinate between states A and B. Then, the free energy profile along the reaction coordinate can be found using enhanced sampling methods to get good population statistics on the recrossings between A and B. Typically, enhanced sampling methods require either adding restraining potentials to guide the system from one basin to another or require multiple independent simulations near the transition site to get good statistics on the crossing of A and B. For example, the pathway can be sampled by umbrella sampling.[1] Or, the pathway can be divided into small bins, each of which has a different biasing potential, allowing for a more accurate reconstruction of the free energy profile using the weighted histogram analysis method (WHAM)[2] or the Multistate Bennett Acceptance Ratio (MBAR).[3,4]

Current approaches assume that a pathway has a dominant route and some small ensemble of variations around it. For example, the *nudged elastic band* method[5,6] starts with an assumed dominant path but allows for deviations from it by spring-law forces.[7] *Metadynamics*[8] assumes a pathway is defined by a set of collective variables (CV) and then samples those paths efficiently using history-dependent biases along them to force the sampling into regions not sampled before. This methodology is often combined with parallel tempering for more efficient sampling.[9,10] Here, big challenges include guessing good CVs and the hysteresis and inaccuracies that arise when the CVs do not reflect well the underlying pathways.[9]

Some current methods use multiple independent runs rather than biasing potentials to overcome barriers and sampling limitations. In *transition path sampling*[11,12] many independent trajectories are started from states near a possible transition path in order to collect statistics on which paths traverse from one basin to another. The approach is highly parallelizable and the computational cost is linear in the barrier (W_{AB}), whereas it is exponential for a direct (MD) approach (whenever $W_{AB} \gg$

Received: December 28, 2017
Published: March 16, 2018

DOI: 10.1021/acs.jctc.7b01294
J. Chem. Theory Comput. 2018, 14, 2109−2116

k_BT). In *milestoning*,[13] the energy surface is broken into different sections, and trajectories are initiated in each to record crossings between sections. This allows us to reconstruct the whole kinetic landscape. In the *weighted ensemble path sampling*[14,15] method, different trajectories are started from the same point (or bin), each carrying a weight of 1/M where M is the number of runs started. After each simulation, statistical weights for bins are calculated, and more simulations are started from the end points. By keeping track of the different weights and resampling, the kinetics and thermodynamics can be efficiently recovered at the end of the simulation. Recently, Markov State Models (MSM) have become a popular way[16−20] to describe the kinetics as memoryless jumps between metastable states. Transition probabilities between states are estimated from either long or many short independent MD trajectories. Also, methods to build MSMs from independent trajectories include adaptive sampling[21−23] (spawning simulations in regions of higher uncertainties) and adaptive seeding[24,25] for choosing good starting configurations (e.g., the FAST algorithm[26]) for spawning short trajectories.

In summary, here are the challenges. First, good reaction coordinates or CVs are rarely known in advance. Second, for computational practicality, it is often assumed that only one reaction path dominates. Yet, in important problems like protein folding, the route structure is highly diverse and heterogeneous. Third, obtaining the free energies, barriers, and kinetics along paths is computationally expensive because it requires expensive enhanced sampling of closely spaced probability distributions stepping along the whole path. We describe below MELD-path, which first seeks relevant states on the whole conformational surface by MELD-accelerated MD and then finds free energies and kinetics by seeding unbiased trajectories from these states. Below, we first review the MELD method for accelerating the MD searching for relevant states, given the two end states A and B.

MELD Samples States and Populations. MELD (Modeling Employing Limited Data) is a method that accelerates MD simulations when at least some information is known.[27,28] Using Bayesian modeling, MELD "melds together" physical simulations, such as MD with force fields, with external information on some kind that need not be well conditioned. MELD-accelerated MD preserves detailed balance. Hence, populations are relevant and related to free energies using Boltzmann weights. MELD uses a Hamiltonian and temperature replica exchange protocol where the Hamiltonian is modified with biasing potentials to satisfy general knowledge.[27,28] It has been validated for folding small proteins,[28,29] in protein structure determination,[27] and in finding the binding poses and affinities of peptides binding to proteins.[30,31] MELD can speed up in sampling rare events; for example, NuG2 (a designed fast folding variant of protein G)[32] can be folded starting from a completely extended chain within 500 ns of MELD simulations and detected as the lowest-free-energy cluster,[28] whereas unaccelerated MD simulations exceeding 50 μs do not sample the native state.[33] So far, MELD has only been proven as a method for finding stable or metastable conformational states of proteins, not of mechanistic pathways. In the present work, we show how MELD can be used to identify metastable states and heterogeneous reaction coordinates and give populations and rates.

Modeling the Helix-to-Helix Transition of AIB. As a proof-of-principle, we test MELD-path on the Aib_9 peptide (Figure 1). This is a good test system because (1) extensive unaccelerated MD simulations are already available. (2) AIB exhibits hierarchical dynamics and is too complex to sample efficiently via long MD (the unaccelerated simulations are not fully converged). (3) The full set of metastable states are readily enumerated. (4) It is symmetric (because AIB, α-aminoisobutyric acid, is achiral; see SI Figure 1), providing for a strong internal check on accuracy and convergence (of states, kinetics and pathways). (5) From the nature of the possible states, it is clear that many different routes between the end states are possible.

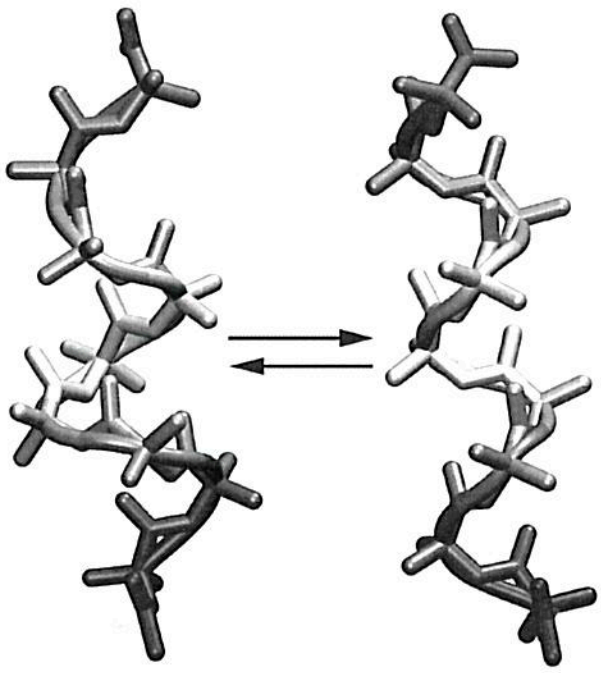


Figure 1. Left- to right-helix transition of AIB9 studied here.

Despite it is short length, the Aib_9 peptide forms very stable 3_{10} helices[34−37] and is able to stabilized shorter living α-helices. This system has been broadly studied both computationally and experimentally for energy transport along its chain,[38,39] giving good qualitative results between the two and exhibiting a dynamical transition behavior that is well reproduced computationally. Long MD trajectories carried out previously[40] show two kinds of events are needed for transitions between left (l) and right (r) helix conversion (which happens in the microsecond time scale): (1) hydrogen bond transitions in the picosecond time scale and (2) transitions of individual residues (l/r) at the nanosecond time scale. Thus, in MD simulations, many hydrogen bond transitions and conformational switching of individual residues are observed but few complete helix-to-helix transitions. Standard MD does not obtain converged populations and kinetics of the system, as judged with the imperfect symmetry in time scales and populations.[41] Here, we focus on the behavior of the central five residues of this peptide in order to avoid end effects. Each residue can be classified into three dominant states of the Ramachandran map: left helix region (l), right helix region (r), or neither ($-$). Hence, there are $3^5 = 243$ possible states for this system.

DOI: 10.1021/acs.jctc.7b01294
J. Chem. Theory Comput. 2018, 14, 2109−2116

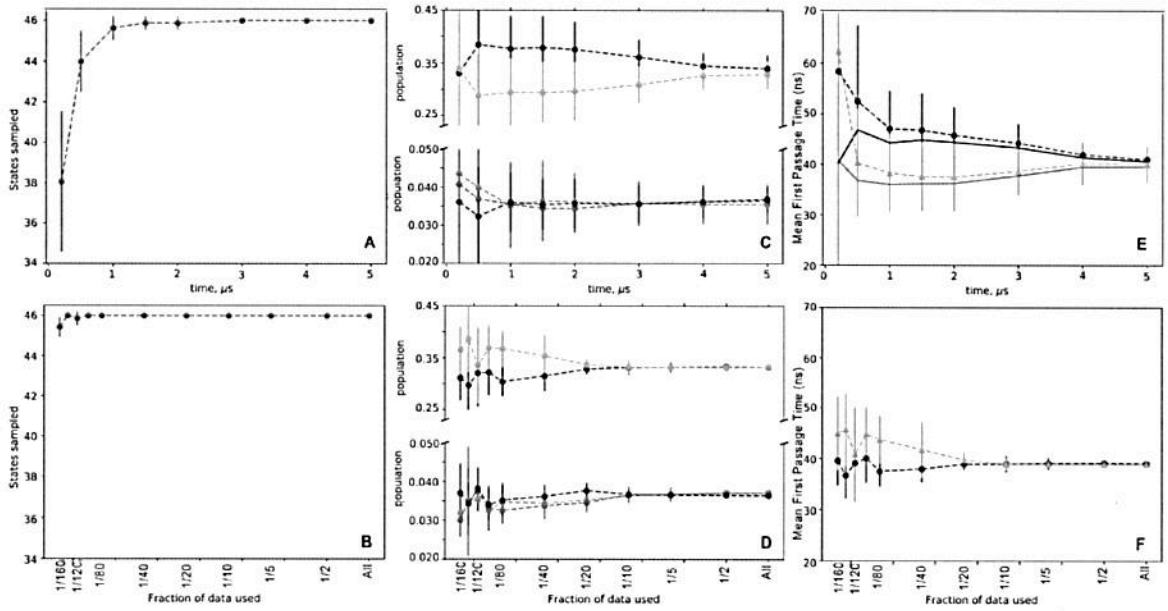


Figure 2. Sampling needed to achieve convergence with long MD or the MELD-path approach. (Top row) Single MD trajectories as a function of simulation time, showing averages and standard deviations over 16 independent runs. (Bottom row) Showing MELD-path quantities obtained using a fraction x = all, 1/2, 1/5,··· of the full run to construct the MSMs. (A and D) Amount of sampling needed to visit all 46 relevant states. (B and E) Convergence of the top five state populations. (C and F) Convergence of the MFPT for going from left- to right-handed helix or vice versa. The red and black lines in panel C represent using all 16 independent simulations to construct the MSM.

2. METHODS

Details of the Present Simulations. We used AIB parameters derived for the NCAA[42] (Non Canonical Amino Acid force field library) compatible with the AMBER molecular package[43] using the GBNeck2 implicit-solvent model.[33] For the present problem, the force field and implicit solvent model are able to reproduce well the helical conformations of the molecule. The kinetics predicted within the implicit solvent model are likely to be accelerated relative to explicit solvent or experiments. The MELD-path protocol discussed here can also be applied using explicit solvent at a higher computational expense, but the point of the present work is just to test the sampling approach in MELD-path.

MELD Runs. The MELD[27,28] plugin to OpenMM[44] is used to run the H,T-REMD. The innovation comes in the way the Hamiltonian is changed: we input information about the system based on the knowledge that Aib_9 likes to make helices, with the understanding that some of the information will be accurate and some will not. During MELD trajectories, only the data that are most compatible with the current conformation are used until the next time step. The data are enforced with flat bottom harmonic restraints that vanish at the higher replicas and become stronger at the lower ones; thus, the system samples from completely unfolded structures to structures that are compatible with a part of the data and the force field at lower replicas. We enforced three different types of protocols to assess whether our results were independent of the information we used (three independent MELD runs).

Here, we describe one of the protocols. The other two protocols are described in the SI. Since the helices can be either 3−10 or α-helices, we input all possible hydrogen bond patterns O(i) → N(i + 3) or O(i) → N(i + 4) to be within 4 Å of each other as flat bottom harmonic restraints. But, we only ask that one restraint be satisfied at any point during the trajectory. Note that this resembles an experiment that mimics low-quality NMR NOE data. Indeed, this is what MELD has successfully been shown to do:[29−31] to handle sparse, noisy, and ambiguous data.[27,28,45] Consequently, MELD-path is an extensible approach to more complex problems of folding and binding than the small peptide conformational transition described here. The use of data accelerates the nucleation of helical states. Since the system is symmetric, these restraints do not favor either left- or right-handed helices. We ran MELD with 30 replicas for 2.5 μs, requiring under a week of computation on our local GPU cluster.

MELD-Path. MELD by itself is a method for searching over states, not for giving kinetics. In contrast, MELD-path uses unbiased simulations seeded from states found by MELD to recover both kinetics and state populations. Relevant states were chosen as seeds for generating unbiased simulations. We use three seed structures for each state originating from the three protocols described above and in the SI. For each seed, we run 20 independent simulations, each of which runs for about 15 ns in implicit solvent (25 min computer time limit on a single GPU in the Blue Waters supercomputer) using the Amber MD package.[43] This corresponds to 13,805 simulations for an aggregated simulation time of about 221 μs. The 25 min time limit takes advantage of the backlogging in the queueing system, allowing us to collect 221 μs of unbiased MD simulations in 17 days (a sequential run of the same length would have taken 255 days if it could run continuously; our approach would have taken under 2 h if we could use all Blue Waters GPU nodes simultaneously).

Long, unbiased MD. We ran 16 independent simulations with the same implicit solvent and force field parameters as the MELD and MELD-path runs. Each trajectory was at least least 5 μs, summing to a total of 80 μs. Each independent trajectory

DOI: 10.1021/acs.jctc.7b01294
J. Chem. Theory Comput. 2018, 14, 2109−2116

took about 6 days of computer time using the same resources as above.

3. RESULTS AND DISCUSSION

MELD Samples the Relevant States and Populations Well. Out of the possible 243 states, the initial MELD simulations samples 231 states (the other two protocols described here produce 229 and 228 states). SI Figure 2 shows good symmetry, and evaluation of all the states across the three different protocols shows good agreement. However, this characterization of states is not optimal; many states are kinetically indistinguishable and interconvert rapidly (SI Figure 2). Hence, following previous work,[46] we processed the ensemble of trajectories using standard Markov state modeling software[21,46,47] as well as our own approach.[48,49] We featurized MELD trajectories in terms of the phi and psi dihedrals of each internal residue and projected them onto the principal components (tICA and dPCA+,[49] see SI Figure 3 for projections of principal components). Both tICA and dPCA+ analyses of the unaccelerated MD and the MELD simulations show a better definition of states based on kinetic clustering despite the limitation that in our MELD replica exchange approach kinetics are biased. dPCA+ categorized residues to be in one of four states: *l*, *r*, *l**, or *r**, where ∗ defines excited states (SI. Figure 1). According to this definition, there are 46 macrostates. We use this definition from now on.

MELD correctly identifies the two helical states as the lowest in free energy, and the slowest conformational transition in the system corresponds to the helix-to-helix transition (SI Figure 2). There are an additional 20 metastable states in the system, none involving excited states (which have lower populations). These results are in agreement with previously published[41] long unbiased simulations, but they are obtained in fraction of the time. The 2.5 μs required 5 days of sampling, while unbiased simulations of the same length are much less efficient at sampling the conformational landscape. The caveat is that MELD kinetics are biased since MELD uses replica exchange, but they can still capture the underlying routes. The MELD-path protocol uses these states as a starting point to produced corrected kinetics. However, before describing MELD-path results, it is worth noting similarities and differences with unbiased simulations. Rather than using published data, we started several long, independent unbiased simulations (see Methods) from the same configuration. This is to ensure a fair comparison *vis-a-vis* the solvent model, capping groups, and force field parameters. We did independent simulations in order to get error estimates and compare overall performance with MELD-path.

The dPCA+ decomposition exhibits a qualitatively similar free energy surface between MELD simulations and the pooling of the 16 long MD simulations. In all cases, the results are consistent with 46 main states. It takes a bit over 2 μs of sampling with unbiased MD to ensure visiting all 46 states (Figure 2). MELD runs cover the 46 states in the range from 0.3 to 0.6 μs in the three protocols we tested, a significant speed up (less than a day of sampling in our local cluster). However, MELD does not provide kinetics and hence only provides good starting structures for the MELD-path approach.

MELD-Path Samples Well the Kinetic Routes. We used MELD-path starting from MELD states to sample 221 μs of unbiased MD as described in Methods. The Ramachandran plot is identical for all five internal residues analyzed here (SI Figure 1). The resulting dPCA+ plot is perfectly symmetrical, with the left-hand helix and right-hand helix being the two most populated states (Table 1 and Figure 3). Clustering using a density-based approach[48] on the subspace of the first five PCA modes, we can again identify 46 microstates (SI Table 1 and Figure 3). At this point, the data coming from pooling the 16 long MD (80 μs of aggregate sampling) or the MELD-path data (221 μs of aggregate sampling) tells us the same information: (1) Both sample 46 states efficiently. (2) The populations of the top states converge to roughly 33% each for the left- and right-handed helix. (3) The mean first passage time for the helix-to-helix transition is of the order of 40 ns. We can investigate how much data is needed to converge to these results. For the long MD, we can estimate Markov models based on shorter chunks of simulation and estimate average and standard deviations from the 16 independent runs, whereas for the MELD-path data we can choose to use a different number of trajectories to construct the Markov state models and choosing different subsets of trajectories estimate averages and deviations. This is summarized in Figure 2 for the three quantities described above: states, populations, and helix-to-helix kinetics. It is easy to see that MELD-path runs efficiently sample all states even for a fraction of the data. This is expected since by construction we are starting from seeded states representing all states in the system. On the other hand, individual MD needs to be longer than 2 μs to ensure sampling of all 46 states (panels A and B in Figure 2). Looking at the convergence of the population for the top five states (panels C and D in Figure 2), we see that we would need only 1/20th of the data (11 μs to sample to converge the top two states, 1 day of sampling), whereas it would take at least 4 μs from an

Table 1. Metastable Populations (in percent) and Right-/Left-Handed Classification with *r*/*l* Denoting the Main States

state	*l*/*r*	pop	state	*l*/*r*	pop
1	*lllll*	30.9	2	*rrrrr*	30.6
3	*rrlll*	4.4	4	*rrrll*	4.4
6	*lllrr*	4.2	5	*llrrr*	4.2
7	*rllll*	3.8	8	*lrrrr*	3.8
10	*llllr*	3.2	9	*rrrrl*	3.3

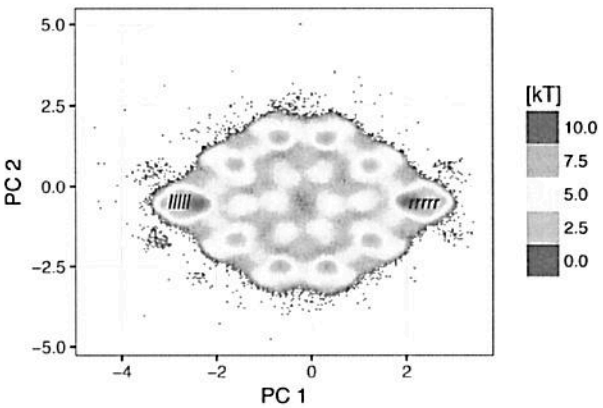


Figure 3. 2D projections of Aib9 of MELD-path trajectories on the two first eigenvectors from dPCA+. Notice that MELD+MD samples the landscape significantly better than just MELD (SI Figure 2), and the symmetry of the molecule is captured in the symmetry in state populations.

DOI: 10.1021/acs.jctc.7b01294
J. Chem. Theory Comput. 2018, 14, 2109−2116

individual trajectory (4.8 days of sampling). These results hold also for converging the mean first passage time for the helix-to-helix transition (panels E and F in Figure 2). Roughly, we see that we need to wait 5 times longer to sample the same phenomena using long simulations. Note that this is a small system, so the scaling benefits should increase for larger systems.

Finally, SI Figure 4 shows us the free energy surface vs the two principal components for different amounts of sampling. It illustrates the 5 fold increase in efficiency of convergence. It also shows the high degree of symmetry in sampling even in the least populated cases.

Kinetic information further allows us to identify the first 22 states representing 98% of the sampled conformations as metastable and correspond to each residues being in a *l* or *r* region of the Ramachandran plot. Their populations follow symmetry quite well. We looked at the transition matrices at different lag times to estimate the stabilities of these states (SI Figure 5). At short lag times (40 ps), all 46 states have a high self-transition probability, denoted by a high probability value along the diagonal in SI Figure 5 (left plot). The first 22 states exhibit very small probability of jumping to other states on this time scale. However, the next 24 states correspond to short-lived conformations with high probability to transition to one of the top 22 states. Moving to longer time scales (1 ns, right-hand panel in SI Figure 5) shows indeed that the self-transition probability of states 23 to 46 vanishes, and most states have a high probability of transitioning to the top 22 states. The diagonal values in SI Figure 5 (right plot) now exhibit very low probability after state 22. Some residues in states 23−46 correspond to excited states[40] (marked *l** or *r** in SI Figure 1). These excited states have been described before as intermediates on the way to *l*/*r* transitions.

We then looked in more detail at the transitions between the main 22 states. Figure 4 shows the metastable states and the transition rates between them as a network of nodes at their average PC1/PC2 positions (see SI Figure 6 for long MD vs MELD-path on the original 46 states). Node sizes are scaled by the logarithm of state population, and edges are scaled by the transition rate. Blue connections depict transitions to the left (along PC1), while orange connections go to the right. The network is based on the transition matrix with τ = 1 ns. Detailed balance has been checked for this matrix, with differences in population probabilities from forward and backward propagation of $\sim 10^{-4}$, i.e., considering numerical errors detailed balance is fulfilled. Figure 4 highlights again the remarkable symmetry of Aib9 in aspects like state populations, *l*/*r* compositions, and transition rates. Note that the most populated metastable states (1−10) lie on the edges of the plot. Kinetics are converged and return the expected behavior of the system. For example, the transition rate between state one (or two) and any other state in which one residue flips conformation is roughly the same; this is expected since the probability of breaking the helix at any one point in the central five residues is roughly the same.

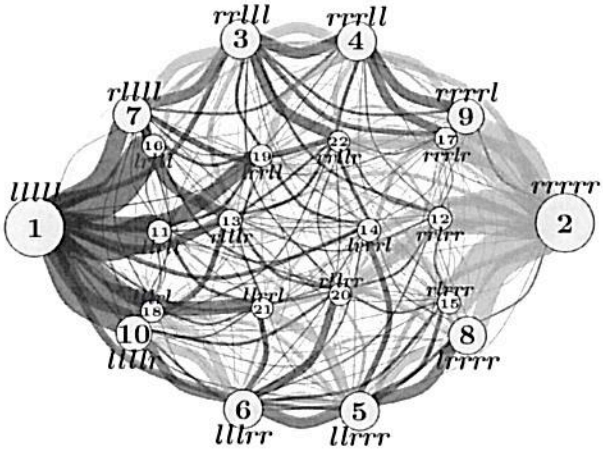


Figure 4. Transition rates between the 22 states and their neighbors from MELD + MSM. MSM representation (τ = 1 ns) of the 22 major states, positioned according to their average of PC1 and PC2. Line thicknesses are proportional to the rates. Blue connections go to the left, orange to the right.

We now quantify the rates and routes by using Markov Chain Monte Carlo sampling (averaging over 10^6 chains) and transition probability propagation. The mean first passage time of the $l \leftrightarrow r$ helix transition has been estimated to be of the order of ~40 ns, compared to ~200 ns from previous results in explicit water. This disagreement in time scales is to be expected as the implicit solvent allows much faster diffusion due to the lower viscosity of the solvent. So, these dynamics are accelerated by a factor of 5, well within the expected range. Further mean first passage times are shown in Table 2. Using

Table 2. Mean First Passage Times (ns) between Important States as Given by Markov Chain Monte Carlo Sampling on MSM with τ = 1 ns

i	*j*	MFPT(*i*,*j*)	MFPT(*j*,*i*)
llllll	*rrrrrr*	40	40
llllll	*llllrr*	41	11
llllll	*lllrrr*	50	20
llllll	*rllllr*	43	15
llllll	*rrllll*	51	23
rrrrrr	*rrrrrl*	41	11
rrrrrr	*rrrlll*	49	20
rrrrrr	*lrrrrr*	44	15
rrrrrr	*llrrrr*	53	23

Bayesian Markov models, we estimate the errors on these transition times for the helix-to-helix transitions to be 40.0 ns (standard deviation of 0.6 ns, standard error of 1.5) and 39.5 ns (standard deviation of 0.6 ns and standard error of 1.6), indicating the robustness and symmetry of the simulated transitions. The transitions toward the (very) high populated all-*l* and all-*r* states are relatively fast (~10−20 ns), and transitions out of these states are about twice as slow, as one would expect. However, regardless of the target state, all transitions out of the stable states happen on about the same time scale, with the all-*l* to all-*r* transitions even being relatively fast in comparison to other transitions.

It may seem paradoxical that a bigger conformational change can happen on the same time scale as the flipping of a single amino acid (Table 2). The explanation for this is the multiplicity of pathways. Upon close inspection of the transition probabilities (Figure 4), it becomes clear that there are multiple pathways transitioning between all-*l* and all-*r* states, whereas the individual paths between specific states are fewer. To illustrate this, we further coarse grain our kinetic model by lumping together states which have the same amount of

DOI: 10.1021/acs.jctc.7b01294
J. Chem. Theory Comput. 2018, 14, 2109−2116

residues in *l*/*r* conformations (e.g., *llrrr* and *lrrrl* are now *l2r3*, see Figure 5). Hence, in SI Table 2, we look at what is the mean

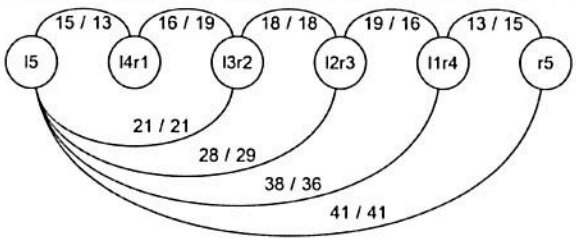


Figure 5. Mean first passage times (in ns) in a coarse-grained representation of the system. Switching one residue at a time takes around 15 ns on average, whereas jumping to any particular microstate inside those coarse-grained states can take much longer. Transitions with multiple residues changing state are also possible. The pairs of numbers refer to forward (i.e., left to right, first number) and backward (right to left, second number) passage times.

first passage time between states with different numbers of *l* or *r* states irrespective of their ordering. As expected, now the probability of going *l5* ↔ *l4r1* ↔ *l3r2* ↔ *l2r3* ↔ *l1r4* ↔ *r5* is close to intuition. It takes about 15 ns for a single residue to flip states, but double or triple jumps are possible, accelerating the process (e.g., *l5* →*l4r1* →*l3r2* takes about 30 ns, whereas a direct jump only takes 22 ns). In the same way, the five independent jumps would take around 75 ns if the amino acids flipped one at a time, but with cooperativity and multiple jumps, we get an average of 40 ns for the transition. Thus, we have shown that the apparent paradox in time scales shown in Table 2 is easily explainable when considering all possible pathways in the system. Although, a transition between states all-*l*/*r* to any particular state is rare, the transition from states all-*l*/*r* to one of the other metastable states is not so rare.

MELD-Path Identifies the Dominant Pathways. We use transition path theory[50,51] as implemented in the PyEMMA software package[47] to calculate the most populated paths between the all-*l* and all-*r* states (see Figure 6 and similar plots comparing long MD and MELD-path on the 46 state system in SI Figure 7). Selecting the six most important pathways for each direction, we show in Figure 6 and in SI Table 3 that most

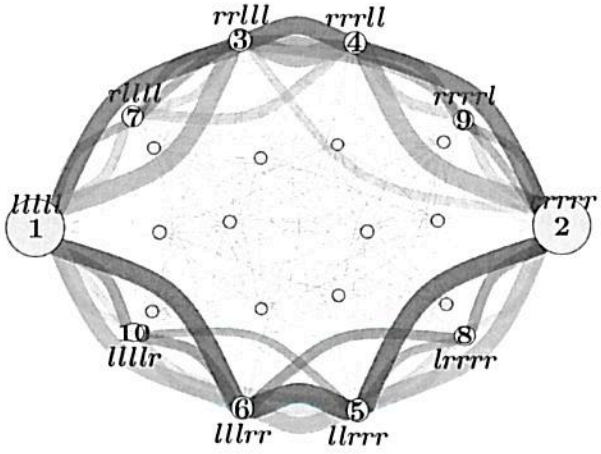


Figure 6. Predicted pathways of maximum net flux from all-*l* to all-*r* (orange) and vice versa (blue) coming from transition path theory.

of the transitions occur when an all-*l* or all-*r* configuration sequentially loses *l* or *r* residues, starting from one end or the other. These pathways alone describe about 60% of the total flux between the all-*l* and all-*r* states (SI Figure 8). Significantly less frequently does the molecule transition directly from all-*l* to all-*r* (or the reverse). While specific pathways with configuration changes in the middle of the chain are also very infrequently sampled, as a total they still describe roughly the other 40% of total flux. The comparison between long MD and MELD-path can also be seen in SI Figure 7.

4. SUMMARY

We describe MELD-path, a computational accelerator for molecular dynamics simulations that finds reaction pathways between conformational states of biomolecules. It broadly samples kinetically relevant states and is correspondingly able to find broad ensembles of routes, if they exist. We give a proof of principle on the Aib_9 peptide helix-to-helix transition. The Markov-State Model seeding is embarrassingly parallelizable: in the limit of enough GPUs, all 13,805 independent trajectories used in this work could be collected in 25 min, whereas a single aggregate trajectory would require 255 days (requiring the same total GPU time). Our 30-replica MELD run took 1 week, and the collection of all the short MD trajectories took 17 days, giving a huge speedup relative to a single trajectory. The method can readily be adapted for explicit solvent. Also, the generation of initial conformations could have been based on geometric sampling of dihedrals at a lower computational cost. In short, we believe the MELD-path method may also be a general and efficient way to explore more complex mechanisms and pathways.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.7b01294.

Supplementary methods, tables, and figures. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: dill@laufercenter.org. Phone: 631-632-5400.

ORCID

Alberto Perez: 0000-0002-5054-5338

Gerhard Stock: 0000-0002-3302-3044

Ken Dill: 0000-0002-2390-2002

Author Contributions

§A. Perez and F. Sittel contributed equally to the work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research is part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (Awards OCI-0725070 and ACI-1238993) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana−Champaign and its National Center for Supercomputing Applications.This work is also part of the PRAC allocation support by the NSF Award ACI1514873. K.D. and A.P. also appreciate the support from NIH Grant GM125813 and from the Laufer Center. G.S. has been

DOI: 10.1021/acs.jctc.7b01294
J. Chem. Theory Comput. 2018, 14, 2109−2116

supported by the Deutsche Forschungsgemeinschaft via Grant STO 247/11.

REFERENCES

(1) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187−199.

(2) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **1992**, *13*, 1011−1021.

(3) Bennett, C. H. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **1976**, *22*, 245−268.

(4) Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **2008**, *129*, 124105.

(5) Elber, R.; Karplus, M. A method for determining reaction paths in large molecules: Application to myoglobin. *Chem. Phys. Lett.* **1987**, *139*, 375−380.

(6) Mills, G.; Jonsson, H. Quantum and thermal effects in H2 dissociative adsorption: Evaluation of free energy barriers in multidimensional quantum systems. *Phys. Rev. Lett.* **1994**, *72*, 1124−1127.

(7) Bergonzo, C.; Campbell, A. J.; Walker, R. C.; Simmerling, C. A partial nudged elastic band implementation for use with large or explicitly solvated systems. *Int. J. Quantum Chem.* **2009**, *109*, 3781−3790.

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

(9) Bussi, G.; Gervasio, F. L.; Laio, A.; Parrinello, M. Free-energy landscape for beta hairpin folding from combined parallel tempering and metadynamics. *J. Am. Chem. Soc.* **2006**, *128*, 13435−13441.

(10) Barducci, A.; Bussi, G.; Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **2008**, *100*, 020603.

(11) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. Transition Path Sampling: Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291−318.

(12) Chandler, D. Statistical mechanics of isomerization dynamics in liquids and the transition state approximation. *J. Chem. Phys.* **1978**, *68*, 2959.

(13) Faradjian, A. K.; Elber, R. Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.* **2004**, *120*, 10880−10889.

(14) Huber, G. A.; Kim, S. Weighted-ensemble Brownian dynamics simulations for protein association reactions. *Biophys. J.* **1996**, *70*, 97−110.

(15) Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. The "weighted ensemble" path sampling method is statistically exact for a broad class of stochastic processes and binning procedures. *J. Chem. Phys.* **2010**, *132*, 054107.

(16) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **2010**, *52*, 99−105.

(17) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. A.; Pande, V. S. Markov State Model Reveals Folding and Functional Dynamics in Ultra-Long MD Trajectories. *J. Am. Chem. Soc.* **2011**, *133*, 18413−18419.

(18) Noé, F. Probability distributions of molecular observables computed from Markov models. *J. Chem. Phys.* **2008**, *128*, 244103.

(19) Noé, F.; Schuette, 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*, 19011−19016.

(20) Stanley, N.; Esteban-Martín, S.; De Fabritiis, G. Kinetic modulation of a disordered protein domain by phosphorylation. *Nat. Commun.* **2014**, *5*, 5272.

(21) Doerr, S.; Harvey, M. J.; Noé, F.; De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.* **2016**, *12*, 1845−1852.

(22) Plattner, N.; Doerr, S.; De Fabritiis, G.; Noé, F. Complete protein-protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nat. Chem.* **2017**, *9*, 1005−1011.

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

(24) 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*, 19765−19769.

(25) Biswas, M.; Lickert, B.; Stock, G. Metadynamics Enhanced Markov Modeling of Protein Dynamics. *J. Phys. Chem. B* **2018**, DOI: 10.1021/acs.jpcb.7b11800.

(26) Zimmerman, M. I.; Bowman, G. R. FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs. *J. Chem. Theory Comput.* **2015**, *11*, 5747−5757.

(27) MacCallum, J. L.; Perez, A.; Dill, K. 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.

(28) Perez, A.; MacCallum, J. L.; Dill, K. Accelerating molecular simulations of proteins using Bayesian inference on weak information. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 11846−11851.

(29) Perez, A.; Morrone, J. A.; Brini, E.; MacCallum, J. L.; Dill, K. Blind protein structure prediction using accelerated free-energy simulations. *Sci. Adv.* **2016**, *2*, e1601274−e1601274.

(30) Morrone, J. A.; Perez, A.; MacCallum, J.; Dill, K. Computed Binding of Peptides to Proteins with MELD-Accelerated Molecular Dynamics. *J. Chem. Theory Comput.* **2017**, *13*, 870−876.

(31) Morrone, J. A.; Perez, A.; Deng, Q.; Ha, S. N.; Holloway, M. K.; Sawyer, T. K.; Sherborne, B. S.; Brown, F. K.; Dill, K. Molecular Simulations Identify Binding Poses and Approximate Affinities of Stapled α-Helical Peptides to MDM2 and MDMX. *J. Chem. Theory Comput.* **2017**, *13*, 863−869.

(32) Nauli, S.; Kuhlman, B.; Baker, D. Computer-based redesign of a protein folding pathway. *Nat. Struct. Biol.* **2001**, *8*, 602−605.

(33) Nguyen, H.; Roe, D. R.; Simmerling, C. Improved Generalized Born Solvent Model Parameters for Protein Simulations. *J. Chem. Theory Comput.* **2013**, *9*, 2020−2034.

(34) Toniolo, C.; Bonora, G. M.; Benedetti, E.; Bavoso, A.; Di Blasio, B.; Pavone, V.; Pedone, C. Linear oligopeptides: peptaibol antibiotics — preferred conformation of the 2−9 segment of emerimicins III and IV and all related short sequences. *Int. J. Biol. Macromol.* **1985**, *7*, 357−362.

(35) Benedetti, E.; Bavoso, A.; Diblasio, B.; Pavone, V.; Pedone, C.; Crisma, M.; Bonora, G. M.; Toniolo, C. Linear Oligopeptides. 81. Solid-State and Solution Conformation of Homooligo(Alpha-Amino-isobutyric Acids) From Tripeptide to Pentapeptide - Evidence for a 310 Helix. *J. Am. Chem. Soc.* **1982**, *104*, 2437−2444.

(36) Karle, I. L.; Balaram, P. Structural characteristics of alpha-helical peptide molecules containing Aib residues. *Biochemistry* **1990**, *29*, 6747−6756.

(37) Toniolo, C.; Crisma, M.; Formaggio, F.; Peggion, C. Control of peptide conformation by the Thorpe-Ingold effect (C alpha-tetrasubstitution). *Biopolymers* **2001**, *60*, 396−419.

(38) Botan, V.; Backus, E. H. G.; Pfister, R.; Moretto, A.; Crisma, M.; Toniolo, C.; Nguyen, P. H.; Stock, G.; Hamm, P. Energy transport in peptide helices. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 12749−12754.

(39) Nguyen, P. H.; Park, S.-M.; Stock, G. Nonequilibrium molecular dynamics simulation of the energy transport through a peptide helix. *J. Chem. Phys.* **2010**, *132*, 025102.

(40) Buchenberg, S.; Schaudinnus, N.; Stock, G. Hierarchical Biomolecular Dynamics: Picosecond Hydrogen Bonding Regulates Microsecond Conformational Transitions. *J. Chem. Theory Comput.* **2015**, *11*, 1330−1336.

(41) Buchenberg, S. Energy and Signal Transport in Proteins: A Molecular Dynamics Simulation Study. Ph.D. Thesis, freidok.uni-freiburg.de, University of Freiburg, 2016.

DOI: 10.1021/acs.jctc.7b01294
J. Chem. Theory Comput. 2018, 14, 2109−2116

(42) Khoury, G. A.; Smadbeck, J.; Tamamis, P.; Vandris, A. C.; Kieslich, C. A.; Floudas, C. A. Forcefield-NCAA: ab initio charge parameters to aid in the discovery and design of therapeutic proteins and peptides with unnatural amino acids and their application to complement inhibitors of the compstatin family. *ACS Synth. Biol.* **2014**, *3*, 855−869.
(43) Case, D. A.; Darden, T. A.; Iii, T. E. C.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Götz, A. W.; Kolossváry, I. W.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Salomon-Ferrer, R.; Sagui, C. A.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A.; Cheatham, T. E.; Goetz, A. W.; Kolossvai, I. *AMBER12*; University of California: San Francisco, 2012.
(44) Eastman, P.; Friedrichs, M. S.; Chodera, J. D.; Radmer, R. J.; Bruns, C. M.; Ku, J. P.; Beauchamp, K. A.; Lane, T. J.; Wang, L.-P.; Shukla, D.; Tye, T.; Houston, M.; Stich, T.; Klein, C.; Shirts, M. R.; Pande, V. S. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J. Chem. Theory Comput.* **2013**, *9*, 461−469.
(45) Pilla, K. B.; Gaalswyk, K.; MacCallum, J. L. Molecular modeling of biomolecules by paramagnetic NMR and computational hybrid methods. *Biochim. Biophys. Acta, Proteins Proteomics* **2017**, *1865*, 1654−1663.
(46) 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*, 3412−3419.
(47) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, *11*, 5525−5542.
(48) Sittel, F.; Stock, G. Robust Density-Based Clustering To Identify Metastable Conformational States of Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 2426−2435.
(49) Sittel, F.; Filk, T.; Stock, G. Principal component analysis on a torus: Theory and application to protein dynamics. *J. Chem. Phys.* **2017**, *147*, 244101.
(50) Weinan, E.; Vanden-Eijnden, E. Towards a Theory of Transition Paths. *J. Stat. Phys.* **2006**, *123*, 503.
(51) 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*, 19011−19016.

DOI: 10.1021/acs.jctc.7b01294
J. Chem. Theory Comput. 2018, 14, 2109−2116