Title
Long-time protein folding dynamics from short-time molecular dynamics simulations

Permalink
https://escholarship.org/uc/item/92q5z171

Journal
Multiscale Modeling & Simulation, 5(4)

ISSN
1540-3459

Authors
Chodera, J D
Swope, W C
Pitera, J W
et al.

Publication Date
2006

Peer reviewed

LONG-TIME PROTEIN FOLDING DYNAMICS FROM SHORT-TIME MOLECULAR DYNAMICS SIMULATIONS[*]

JOHN D. CHODERA[†], WILLIAM C. SWOPE[‡], JED W. PITERA[‡], AND KEN A. DILL[§]

Abstract. Protein folding involves physical timescales—microseconds to seconds—that are too long to be studied directly by straightforward molecular dynamics simulation, where the fundamental timestep is constrained to femtoseconds. Here we show how the long-time statistical dynamics of a simple solvated biomolecular system can be well described by a discrete-state Markov chain model constructed from trajectories that are an order of magnitude shorter than the longest relaxation times of the system. This suggests that such models, appropriately constructed from short molecular dynamics simulations, may have utility in the study of long-time conformational dynamics.

Key words. Markov chain model, molecular dynamics, peptide dynamics, protein folding

AMS subject classifications. 60J10, 60J20, 80A30, 92C05, 92C45, 70K70, 74A25

DOI. 10.1137/06065146X

1. Introduction. Proteins can fold to well-defined native topologies[1] with surprising determinism. Many small, single domain proteins can fold rapidly, reversibly, cooperatively, and without the aid of other molecular machinery. In response to an environmental perturbation such as the introduction or removal of denaturant or a rapid change in solvent temperature, these fast-folding proteins exhibit nearly exponential relaxation kinetics with observed time constants on the order of microseconds. Other proteins exhibit slow and complex kinetics, suggesting the presence of one or more kinetic intermediates. A detailed understanding of this process has been the focus of much of modern biophysics. Ultimately, knowledge of the general mechanistic features by which proteins fold and aggregate is critical for understanding a variety of folding and misfolding diseases, elucidating principles necessary for effective protein design, and developing the basic tools needed for other related technological applications of complex molecular structures.

A description of the mechanism by which a particular protein folds must by necessity be a statistical one. While the initial microscopic state[2] and dynamical trajectory may differ for each molecule in an experiment, many proteins refold to their native (folded) topologies upon the restoration of native conditions with the certainty of macroscopic law [2]. A proper statistical description would summarize the salient features and relative probabilities of relevant folding routes in a way that

[1]By *native topology* or *native structure*, we refer to the *ensemble* of configurations sharing a coarse overall structure, or fold, with the experimental structure.

[2]By *microscopic state*, we refer to the set of generalized coordinates and momenta that completely determine the microscopic state of the system, such as the phase space point.

is meaningful to the physical chemist. This manner of model has been difficult to extract from experiments. Despite the high time resolution possible with optical spectroscopy, the majority of these experiments rely on the observation of an *ensemble* of molecules to obtain sufficient signal, resulting in the ability to observe only (possibly time-dependent) ensemble averages, rather than the behavior of any single molecule. While observations of single molecules are now possible with fluorescence techniques, atomic force microscopy, or optical traps, high temporal resolution is sacrificed to achieve sufficient signal for reliable measurement. In contrast, computer simulation promises the ability to produce information with both atomic detail and high temporal resolution.

In practice, however, the presence of fast vibrational motion constrains the fundamental integration timestep to femtoseconds in order to ensure stability, limiting practical straightforward molecular dynamics simulations of atomically detailed representations of solvated proteins to tens of nanoseconds. As even the fastest folding proteins exhibit relaxation timescales of several microseconds [24], this leads to a *timescale gap* of at least two decades in time. Using supercomputers such as Blue Gene [18] and software specialized for molecular dynamics simulations on these computer systems [17, 19], one can produce atomistic simulations of protein molecules with explicit representation of surrounding solvent on several microsecond timescales. However, the number of trajectories that need be generated to provide an adequate *statistical* characterization of the folding mechanism of even a single protein makes such an endeavor extremely challenging. Distributed computing projects such as Folding@Home [36] regularly collect tens of thousands of trajectories tens of nanoseconds in length, but extracting insight about microsecond timescale dynamics from these large datasets can be difficult [16, 35, 30].

Kinetics models may provide the necessary link between short simulations of a single molecule and long experimental observations of ensembles of molecules. If time evolution of a protein system is characterized by long waiting times within metastable states punctuated by infrequent transitions between these states, interstate dynamics may appear stochastic and memoryless on some short timescale. In this case, long trajectories may be modeled as a Markov chain realized on a discrete state space of a (hopefully small) number of states. While this model could not describe dynamical behavior at very short timescales, which is dominated by molecular motion *within* a metastable state, it could nevertheless faithfully describe long-timescale transitions *between* states. This approach would have numerous advantages. It is precisely these slow transitions involving major structural rearrangements that are of primary interest; elimination of high-frequency detail is often desirable in aiding interpretation of trajectories. To generate a statistical description of folding dynamics, instead of generating many simulations, each long enough to contain complete folding events, we need only generate simulations long enough to characterize transition rates between pairs of conformational substates. Construction would therefore be amenable to parallelization on loosely coupled grids of computer systems. The resulting kinetic model could then be used to compute the stochastic temporal evolution of either a single molecule or an ensemble of molecules, allowing direct comparison to data from both kinds of kinetics experiments, or to answer statistical questions about folding pathways and mechanisms that are currently experimentally inaccessible.

This proposition is not entirely novel. Several groups have constructed stochastic kinetic models from states defined by local potential energy minima of small peptides, using transition state theory to estimate interstate transition rates [9, 26, 3, 28, 33, 32]. Unfortunately, the number of minima grows rapidly with increasing system size, mak-

ing the procedure prohibitively expensive for larger proteins or systems containing explicit solvent molecules. Other work [11, 50, 46, 1, 47, 40] has focused on the construction of discrete- or continuous-time Markovian models to describe dynamics between a small number of states. These models, however, have yet to demonstrate that they can adequately describe the dynamics on timescales much longer than the trajectories from which the models were constructed; no attempt is made to compare the dynamics predicted by the model with long trajectories of explicitly solvated systems. Transition interface sampling [31] and "milestoning" [15] attempt to describe dynamics along a one-dimensional reaction coordinate, but these approaches are valid only if it can be shown that relaxation transverse to this coordinate happens very quickly. Some have suggested constructing stochastic models of dynamics by expansion of the dynamical operator in a smooth basis set [43, 51, 44], but this approach unfortunately suffers from the great difficulty of choosing rapidly convergent basis sets for large molecules. In spite of the challenges with their construction, Markov models that can accurately capture the long-time kinetics of a system can be very useful. Such models embody a concise description of the various kinetic pathways and their relative likelihood. Moreover, they can be used to facilitate the computation of useful properties such as state lifetimes [49], mean first-passage times [46], the existence of hidden intermediates [34], and P_{fold} values (transmission coefficients) [27].

Here we present a proof of principle for how the dynamics of a solvated biomolecular system can be described using information from short simulations. This is illustrated using terminally blocked alanine, a system small enough that its dynamical behavior can also be thoroughly characterized by straightforward molecular dynamics simulation. First, a parallel tempering simulation is conducted to explore the thermally relevant regions of configuration space. Next, a set of metastable states, corresponding to regions of configuration space with low probabilities of escape, are identified. Due to the simplicity of the system considered in this work, these states can readily be identified by hand. This removes the complication of choice of state decomposition for arbitrary macromolecular systems, which we shall not address here. Finally, a number of short trajectories are initiated from each state, and a Markov chain model is constructed from analysis of the observed interstate transitions. We demonstrate the validity of the model by comparing its prediction for the long-time evolution of a nonequilibrium ensemble with what is actually exhibited by an ensemble of long simulations.

This paper is organized as follows. In section 2, the Markov chain model and its method of construction are described. In section 3, the method is applied to terminally blocked alanine in explicit solvent. A discussion of the significance of this result, as well as problems remaining to be solved before the method can be applied to larger biomolecules, follows in section 4.

2. Theory.

2.1. Conformational dynamics as a Markov process. Consider the dynamics of a macromolecular system in equilibrium at some temperature of interest, where we have decomposed all configuration space into a set of M disjoint but contiguous states. If we observe a trajectory of this system at times $t = 0, \tau, 2\tau, \ldots, n\tau$, where τ denotes the observation interval, we can represent the trajectory in terms of the state the system occupies at each of these discrete times, $s_0, s_1, s_2, \ldots, s_n$. The sequence of states produced by such a trajectory is a *discrete-time stochastic process*. If this process is a Markov chain, it must satisfy the *Markov property*, whereby the probability of observing the system in state s_n at the time point $n\tau$, given the state history

$s_0, s_1, s_2, \ldots, s_{n-1}$, is independent of all but the previous state s_{n-1}. For a stationary process which has no explicit dependence on time, this property is given by

$$(2.1) \qquad P(s_n|s_0, s_1, s_2, \ldots, s_{n-1}) = P(s_n|s_{n-1}).$$

As there are a finite number of states, this process can be entirely characterized by an $M \times M$ *transition matrix* $\mathbf{T}(\tau)$ dependent only on lag time τ. The element $T_{ji}(\tau)$ denotes the probability of observing the system in state j at time τ, given that it was initially in state i at time 0:

$$(2.2) \qquad T_{ji}(\tau) \equiv P(j|i).$$

If we do not know the precise initial state of the system at time 0 but only the probability the system started in each state, or if we observe an ensemble of many noninteracting systems in an experiment, we can instead consider the probability of finding one particular molecule in each state i at time $n\tau$ as components of the vector of state probabilities $\mathbf{p}(n\tau)$. If the initial probability vector is given by $\mathbf{p}(0)$, we can write the probability vector at some later time $n\tau$ as

$$(2.3) \qquad \mathbf{p}(n\tau) = \mathbf{T}(n\tau)\mathbf{p}(0) = [\mathbf{T}(\tau)]^n\mathbf{p}(0).$$

This property is described by the *Chapman–Kolmogorov equation*.

2.2. Construction of the Markov chain model from simulation. For a system in which the dynamical evolution is Newtonian but the initial configurations come from a canonical distribution, Swope, Pitera, and Suits [49] show that the transition probability $T_{ji}(\tau)$ can be written as

$$
\begin{aligned}
T_{ji}(\tau) &\equiv \frac{\langle \chi_j(\mathbf{z}(\tau))\chi_i(\mathbf{z}(0))\rangle}{\langle \chi_i(\mathbf{z}(0))\rangle} \\
&= \frac{\int d\mathbf{z}(0)\, e^{-\beta H(\mathbf{z}(0))}\, \chi_j(\mathbf{z}(\tau))\, \chi_i(\mathbf{z}(0))}{\int d\mathbf{z}(0)\, e^{-\beta H(\mathbf{z}(0))}\, \chi_i(\mathbf{z}(0))} \\
(2.4) \qquad &= \int d\mathbf{z}(0)\, p_i(\mathbf{z}(0))\, \chi_j(\mathbf{z}(\tau)),
\end{aligned}
$$

where $\mathbf{z}$ denotes a point in phase space, $\chi_i(\mathbf{z})$ denotes the indicator function for state i, $\beta \equiv (k_B T)^{-1}$ is the inverse temperature, $H(\mathbf{z})$ is the Hamiltonian, and $p_i(\mathbf{z})$ denotes the canonical distribution restricted to state i:

$$(2.5) \qquad p_i(\mathbf{z}) = \frac{e^{-\beta H(\mathbf{z})}\, \chi_i(\mathbf{z})}{\int d\mathbf{z}\, e^{-\beta H(\mathbf{z})}\, \chi_i(\mathbf{z})}.$$

The final expression of (2.4) simply states the obvious: the transition matrix element $T_{ji}(\tau)$ can be estimated in a straightforward (though potentially inefficient) manner by initiating a number of simulations from configurations selected from a canonical distribution within state i, evolving the dynamics for a time τ, and determining the fraction of trajectories that terminate in state j:

$$(2.6) \qquad T_{ji}(\tau) \approx \frac{N_{ji}(\tau)}{\sum_{j'=1}^{M} N_{j'i}(\tau)}.$$

Here $N_{ji}(\tau)$ denotes the number of trajectories initiated from state i that terminate in state j at time τ. This procedure corresponds to the method proposed earlier by
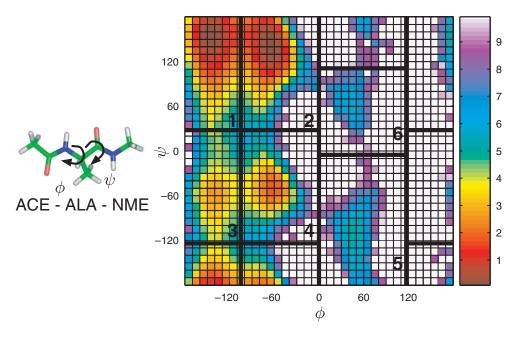
FIG. 1. *Potential of mean force and state boundaries. Left: The terminally blocked alanine peptide with (ϕ, ψ) torsions labeled. Right: The potential of mean force in the (ϕ, ψ) torsions at 302 K estimated from the parallel tempering simulation, truncated at $10k_BT$ (white regions), with reference scale (far right) labeled in units of k_BT. Boundaries defining the six manually identified states are superimposed and the states labeled.*

Swope, Pitera, and Suits in the special case that the *selection cells* from which sets of simulations are initiated are coincident with the states [49].

We do not expect dynamics of a macromolecule in solution to resemble a Markov process for all observation intervals τ, as ballistic motion dominates on very short times, and sufficient time must be allowed for collisions with the solvent and decorrelation of the trajectory within a metastable state. Imperfect definitions of the metastable states may also lead to non-Markovian behavior on short times [49]. At sufficiently long intervals τ, however, we might observe that dynamics resembles a Markov process. While it is impractical to test the condition of complete history independence (see (2.1)), we can simply check the (weaker) condition imposed by the Chapman–Kolmogorov equation (see (2.3)): For transition matrices constructed for a given τ, we check whether (2.3) holds for several lag times $n = 2, 3, 4, \ldots$ to within statistical uncertainty. If so, the Markovian model can be assumed to be a reasonable model of dynamics.

3. Application to terminally blocked alanine peptide.

3.1. System setup and equilibration. Using the `LEaP` program from the AMBER7 molecular mechanics package [6], a terminally blocked alanine peptide (sequence ACE-ALA-NME; see Figure 1) was generated in the extended conformation with peptide force field parameters taken from the AMBER `parm96` parameter set [23]. The system was subsequently solvated with 431 TIP3P water molecules [21] in a truncated octahedral simulation box with dimensions chosen to ensure all box boundaries were at least 7 Å from any atom of the extended peptide. All minimization and

molecular dynamics simulations were conducted using the `sander` program from the AMBER7 package. Default nonbonded cutoffs were used, bonds to hydrogen were constrained with SHAKE using a tolerance of 10^{-8} [39], and long-range electrostatics were treated by the particle-mesh Ewald method [10] with the default settings.

The system was first subjected to 50 steps of steepest descent energy minimization, followed by 1000 steps of conjugate gradient optimization. To equilibrate the explicit solvent system to the appropriate volume, a 100 ps molecular dynamics simulation was performed with the temperature adjusted to 300 K and the pressure to 1 atm by the Berendsen weak-coupling algorithm [4] with temperature and pressure relaxation time constants of 1 ps and 0.2 ps, respectively. The simulation box was fixed at the final size obtained from this equilibration step, with a volume of 13 232 Å^3, in all subsequent simulations.

3.2. Parallel tempering. In order to broadly explore the configuration space of the peptide and ensure that all important conformational substates were located, a parallel tempering (or replica exchange among temperatures) molecular dynamics simulation [48] was conducted using a parallel `Perl` wrapper for the `sander` program [7]. Forty replicas were used, with replica temperatures exponentially distributed over the range 273–600 K, yielding an average exchange acceptance probability of about 50%. All momenta were reassigned from the Maxwell–Boltzmann distribution at the appropriate replica temperature after each exchange attempt, and constant-energy, constant-volume molecular dynamics with a 2 fs timestep was performed between exchange attempts. The algorithm used to select pairs of replicas for temperature exchange attempts starts from the highest temperature replica, attempts to swap the configuration for the next-lowest temperature replica using the Metropolis-like criteria, and proceeds down the temperatures in this manner. On the next iteration, swapping attempts start from the lowest temperature and proceed upward, and this alternation in direction is continued in subsequent pairs of iterations.

Starting all replicas from the volume-equilibrated configuration described above, 100 iterations were conducted with 1 ps between exchange attempts to equilibrate the replicas to their respective temperatures. This equilibration run was followed by a production run of 500 iterations with 20 ps between exchange attempts, a total of 10 ns/replica. Solute configurations and potential energies from the production run were written to disk every 0.1 ps, while full-system restart files were recorded every 1 ps for the purpose of starting new simulations from these configurations, as described in section 3.4.

3.3. State decomposition. The slow degrees of freedom for terminally blocked alanine peptide (neglecting those involving solvent motion) can be captured by the two backbone torsion angles labeled ϕ and ψ (see Figure 1) [5, 29]. To this end, the potential of mean force at 302 K was computed from the parallel tempering data using the weighted histogram analysis method (WHAM) [25, 8] and is shown in Figure 1. Six free energy basins are readily visible, and rectangular regions around these basins were chosen for the decomposition of all of configuration space into a set of six states. State definitions are listed in Table 1 and plotted as thick dividing lines in Figure 1.

3.4. Construction of Markov chain model from short trajectories. To construct a Markov chain model of dynamics once the states were identified, the interstate transition probabilities were computed using the procedure described in section 2.2. A set of 1000 energy-conserving trajectories 10 ps in length were generated from a canonical distribution of initial conditions within each state. This initial

TABLE 1
State definitions for the manual decomposition of (ϕ, ψ)-space into metastable states and populations at 302 K.

State	Label[a]	State definitions		P_{eq}[b]
		ϕ	ψ	
1	C_5	$[117, -105)$	$[28, -124)$	.4787 (.0613)
2	P_{II}	$[-105, 0)$	$[28, -124)$	.4159 (.0486)
3	α_P	$[117, -105)$	$[-124, 28)$	.0425 (.0038)
4	α_R	$[-105, 0)$	$[-124, 28)$	.0588 (.0079)
5	C_7^{ax}	$[0, 117)$	$[111, -5)$	.0030 (.0015)
6	α_L	$[0, 117)$	$[-5, 111)$	.0011 (.0004)

[a]Corresponding state labels from [38].   [b]Equilibrium probabilities at 302 K estimated from the replica exchange simulation by WHAM with corresponding uncertainties representing one standard deviation shown in parenthesis.

distribution was generated by selecting initial configurations from all replicas of the replica exchange simulation with a probability proportional to their weight used in computing canonical averages at 302K, as determined by WHAM, and assigning initial momenta from the Maxwell–Boltzmann distribution. For each lag time τ, an estimate of the transition probability $T_{ij}(\tau)$ was obtained using (2.6). A bootstrap procedure [14], in which 200 replicates of 1000 trajectories from each state were chosen with replacement from the set of trajectories emanating from each state, was used to estimate the uncertainty in the observed transition probabilities.

The observed transition probabilities out of each state as a function of τ are shown in Figure 2, along with the corresponding equilibrium probabilities of each state determined from the replica-exchange simulation. None of the state populations reach their equilibrium values within 10 ps, indicating the slowest relaxation timescales are much longer, perhaps substantially so for trajectories originating from states 5 and 6. Transition matrices at several lag times—0.1 ps, 1 ps, 6 ps, and 10 ps—are shown in Table 2.

3.5. Comparison with long trajectories. To determine the accuracy with which transition matrices constructed from different lag times from short (10 ps) simulations are able to reproduce the statistical dynamics over long times (approximately 100 ps), state populations for an ensemble of trajectories emanating from each state were computed from the model and compared to the observed time evolution of a separate set of long trajectories. For this comparison, 1000 trajectories 100 ps in length were initiated from each state, using the same protocol in section 3.4. Figure 3 shows the time evolution of state populations from these trajectories (along with corresponding uncertainties) as a function of time. Superimposed are state populations computed by (2.3) from the transition matrices constructed for different lag times τ from the short simulations described in section 3.4. These are connected by straight line segments solely to guide the eye; the model cannot make predictions for the populations at times that are not integral multiples of the lag time τ.

The transition probabilities are poorly reproduced in the model constructed with a lag time of only 0.1 ps. Apparently, this time is so short that the system does not behave in a Markovian manner on this timescale. At a lag time of 1 ps, the agreement between the model and long simulations is clearly better, though there are still visible systematic deviations. By a lag time of 6 ps, the agreement is excellent. The model constructed from a lag time of 10 ps also shows excellent agreement, but by this time, the temporal resolution has started to become rather poor. Information about the
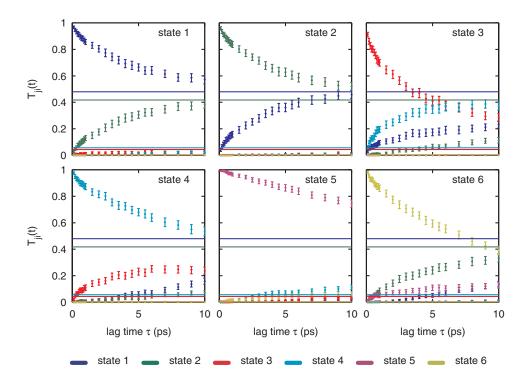
FIG. 2. *Transition matrix elements as a function of lag time estimated from* 10 *ps shooting trajectories. Each plot, labeled above by the state from which the trajectories originated, shows state-to-state transition probabilities as a function of the lag time τ estimated from a set of* 1000 *trajectories* 10 *ps in length originating from an equilibrium distribution within each state. Vertical bars depict* 95% *confidence intervals. Equilibrium state probabilities obtained from the parallel tempering simulations are shown as solid horizontal lines in the corresponding color.*

system is known only for times that are integral multiples of 10 ps. One can imagine that the most useful model would be constructed from the shortest lag time at which dynamics is Markovian, as this model has the highest temporal resolution while still correctly describing long-time dynamics.

3.6. Long-time dynamics from the Markov chain model. As an illustration of the utility of the Markov chain model, Figure 4 depicts an *artificial* trajectory generated by realization of the Markov process, 10 ns in length, three orders of magnitude longer than the short trajectories used to construct the model. While statistical properties of the dynamics can also be extracted in other ways, such as through an eigenvalue decomposition, it may be useful to generate artificial trajectories and analyze them directly. Note the infrequent sampling of states 5 and 6, states with very small equilibrium probabilities, and the long dwell times in the region formed by stable states 1 and 2.

4. Discussion. We have demonstrated that a Markov model constructed from simulations roughly one order of magnitude shorter than the slowest relaxation time in the system is sufficient to capture the long-time dynamics of a simple biomolecular system, terminally blocked alanine peptide in explicit solvent. Instead of generating large numbers of long trajectories to statistically characterize dynamics, we require

<div align="center">
Table 2

Transition matrices[a] at several lag times estimated from a set of 10 ps trajectories.
</div>

$$\mathbf{T}(0.1\text{ ps}) = \begin{bmatrix} .967 & .041 & .029 & & & .002 \\ .030 & .959 & & .003 & & .001 \\ .003 & & .912 & .022 & & \\ & & .059 & .975 & & \\ & & & & .993 & .015 \\ & & & & .007 & .982 \end{bmatrix}$$

$$\mathbf{T}(1\text{ ps}) = \begin{bmatrix} .856 & .161 & .096 & .011 & .002 & .004 \\ .130 & .835 & .008 & .007 & & .086 \\ .014 & .002 & .701 & .109 & .001 & \\ & .002 & .195 & .873 & .014 & \\ & & & & .966 & .047 \\ & & & & .017 & .863 \end{bmatrix}$$

$$\mathbf{T}(6\text{ ps}) = \begin{bmatrix} .642 & .400 & .190 & .068 & .010 & .069 \\ .324 & .586 & .069 & .043 & .011 & .268 \\ .023 & .009 & .373 & .251 & .017 & .002 \\ .011 & .005 & .367 & .637 & .075 & .004 \\ & & .001 & .001 & .839 & .104 \\ & & & & .048 & .553 \end{bmatrix}$$

$$\mathbf{T}(10\text{ ps}) = \begin{bmatrix} .573 & .459 & .232 & .157 & .022 & .138 \\ .385 & .520 & .110 & .072 & .033 & .333 \\ .018 & .013 & .286 & .235 & .030 & .005 \\ .022 & .008 & .371 & .535 & .111 & .009 \\ & & .001 & .001 & .745 & .127 \\ .002 & & & & .059 & .388 \end{bmatrix}$$

[a]Blank entries denote estimated transition probabilities of zero.

only a sufficient number of trajectories to estimate transition probabilities between pairs of states. In addition, these trajectories need only be long enough for interstate dynamics to appear Markovian. Once so constructed, the model can be used to answer various questions of interest regarding the long-time statistical dynamics without the need to perform additional simulations.

While it is impossible to predict what the minimum trajectory length required for Markovian behavior will be for other, larger systems, it is important to recall that most proteins fold on the millisecond to second timescale. Even fast folding proteins can require tens of microseconds to fold [24]. To bring the treatment of these systems within the realm of feasibility, the Markov time would need to remain sufficiently short to allow the collection of a significant number of trajectories despite the presence of relaxation times many orders of magnitude longer. No statement can yet be made about the number of states necessary to model more complex systems or whether this number might make this approach prohibitively expensive.

The question of how best to validate a Markov chain model constructed from short trajectories without additional long-time information is a topic of active research. To determine the lag time to construct the transition matrix so that the Markov chain is an accurate description of long-time dynamics, it was necessary to compare to an additional set of long trajectories. This, of course, defeats the utility of a model constructed from short trajectories. Other methods, such as tests of eigenvalue behavior [49] or direct tests of Markovity [37], may provide alternatives.

In this work, we have avoided the issue of how best to define the number and location of states used for construction of the Markov model. Ideally, these states
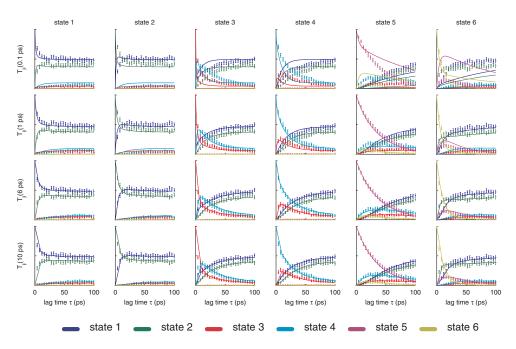
Fig. 3. *Temporal evolution of state populations from Markov chains constructed at different lag times compared with long simulations. Evolution of state probabilities from an ensemble prepared at equilibrium within each state for a Markov model estimated from the set of 10 ps shooting trajectories (solid lines) superimposed on fractional population of each state as a function of time for ensemble of 100 ps trajectories initiated from each state (points). Vertical bars depict 95% confidence intervals in state populations estimated from the long trajectories.*
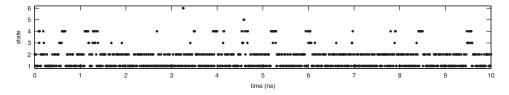

Fig. 4. *An artificial trajectory generated from the transition matrix constructed from a lag time of 10 ps.*

will be significantly *metastable* so that the system rapidly loses memory of its previous location after entering a state, before making a transition to another state. For the system considered, the slow degrees of freedom were known beforehand, so the potential of mean force in these coordinates revealed a useful set of states. In more complex systems, the coordinates in which dynamics is slow will be much more difficult to discern; some automatic method for the identification of metastable states is necessary. Pure conformational clustering methods [22, 20] may prove to be inadequate because they neglect the true locations of kinetic barriers, but attempts to also consider kinetic relationships give promising results but have not yet been applied to large explicitly solvated systems [41, 42, 40]. This problem is the subject of work soon to be reported [45].

Here we employed the most straightforward approach to estimating interstate transition probabilities, whereby a large number of short trajectories are initiated

from equilibrium within each state. While this approach is amenable to distributed or grid computing, the metastable nature of well-chosen states will result in many of these trajectories simply remaining in their state of origin, rather than contributing to estimates of the off-diagonal elements of the transition matrix. It is precisely these off-diagonal elements that are critical in determining which trajectories through state space are most likely. Algorithms employing importance sampling techniques in *trajectory* space—such as transition path sampling [12], transition interface sampling [52], and the string method [13]—may provide an efficient way to compute these interstate transition probabilities.

REFERENCES

[1] M. ANDREC, A. K. FELTS, E. GALLICCHIO, AND R. M. LEVY, *Protein folding pathways from replica exchange simulations and a kinetic network model*, Proc. Natl. Acad. Sci. USA, 102 (2005), pp. 6801–6806.

[2] C. B. ANFINSEN, *Principles that govern folding of protein chains*, Science, 181 (1973), pp. 223–230.

[3] K. D. BALL AND R. S. BERRY, *Realistic master equation modeling of relaxation on complete potential energy surfaces: Kinetic results*, J. Chem. Phys., 109 (1998), pp. 8557–8572.

[4] H. J. C. BERENDSEN, J. P. M. POSTMA, W. F. VAN GUNSTEREN, A. DINOLA, AND J. R. HAAK, *Molecular dynamics with coupling to an external bath*, J. Chem. Phys., 81 (1984), pp. 3684–3690.

[5] P. G. BOLHUIS, C. DELLAGO, AND D. CHANDLER, *Reaction coordinates of biomolecular isomerization*, Proc. Natl. Acad. Sci. USA, 97 (2000), pp. 5877–5882.

[6] D. A. CASE, D. A. PEARLMAN, J. W. CALDWELL, T. E. CHEATHAM, III, J. WANG, W. S. ROSS, C. L. SIMMERLING, T. A. DARDEN, K. M. MERZ, R. V. STANTON, A. L. CHENG, J. J. VINCENT, M. CROWLEY, V. TSUI, H. GOHLKE, R. J. RADMER, Y. DUAN, J. PITERA, I. MASSOVA, G. L. SEIBEL, U. C. SINGH, P. K. WEINER, AND P. A. KOLLMAN, *AMBER7*, University of California, San Francisco, CA, 2002.

[7] J. D. CHODERA, *An MPI-Based Parallel* `perl` *Wrapper to Conduct Replica-Exchange Dynamics for the* `sander` *Molecular Dynamics Program from* `amber7`, http://www.dillgroup.ucsf.edu/~jchodera/code/rex.

[8] J. D. CHODERA, W. C. SWOPE, J. W. PITERA, C. SEOK, AND K. A. DILL, *Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations*, J. Chem. Theor. Comput., to appear.

[9] R. CZERMINSKI AND R. ELBER, *Reaction path study of conformational transitions in flexible systems: Application to peptides*, J. Chem. Phys., 92 (1990), pp. 5580–5601.

[10] T. A. DARDEN, D. M. YORK, AND L. G. PEDERSEN, *Particle mesh Ewald—An N log(N) method for Ewald sums in large systems*, J. Chem. Phys., 98 (1993), pp. 10089–10092.

[11] B. L. DE GROOT, X. DAURA, A. E. MARK, AND H. GRUBMÜLLER, *Essential dynamics of reversible peptide folding: Memory-free conformational dynamics governed by internal hydrogen bonds*, J. Mol. Biol., 309 (2001), pp. 299–313.

[12] C. DELLAGO, P. G. BOLHUIS, AND D. CHANDLER, *Efficient transition path sampling: Application to the Lennard-Jones cluster rearrangements*, J. Chem. Phys., 108 (1998), pp. 9236–9245.

[13] W. E, W. REN, AND E. VANDEN-EIJNDEN, *String method for the study of rare events*, Phys. Rev. B, 66 (2002), 052301.

[14] B. EFRON, *Bootstrap methods: Another look at the jackknife*, Ann. Statist., 7 (1979), pp. 1–26.

[15] A. K. FARADJIAN AND R. ELBER, *Computing time scales from reaction coordinates by milestoning*, J. Chem. Phys., 120 (2004), pp. 10880–10889.

[16] A. R. FERSHT, *On the simulation of protein folding by short time scale molecular dynamics and distributed computing*, Proc. Natl. Acad. Sci. USA, 99 (2002), pp. 14122–14125.

[17] B. G. FITCH, R. S. GERMAIN, M. MENDELL, J. PITERA, M. PITMAN, A. RAYSHUBSKIY, Y. SHAM, F. SUITS, W. SWOPE, T. J. C. WARD, Y. ZHESTKOV, AND R. ZHOU, *Blue Matter, an application framework for molecular simulation on Blue Gene*, J. Parallel Distrib.

Comput., 63 (2003), pp. 759–773.

[18] A. GARA, M. A. BLUMRICH, D. CHEN, G. L.-T. CHIU, P. COTEUS, M. E. GIAMPAPA, R. A. HARING, P. HEIDELBERGER, D. HOENICKE, G. V. KOPCSAY, T. A. LIEBSCH, M. OHMACHT, B. D. STEINMACHER-BUROW, T. TAKKEN, AND P. VRANAS, *Overview of the Blue Gene/L system architecture*, IBM J. Res. Develop., 49 (2005), pp. 195–212.

[19] R. S. GERMAIN, B. FITCH, A. RAYSHUBSKIY, M. ELEFTHERIOU, M. C. PITMAN, F. SUITS, M. GIAMPAPA, AND T. J. C. WARD, *Blue Matter on Blue Gene/L: Massively parallel computation for biomolecular simulation*, in CODES+ISSS '05: Proceedings of the 3rd IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis, ACM, New York, 2005, pp. 207–212.

[20] F. A. HAMPRECHT, C. PETER, X. DAURA, W. THIEL, AND W. F. VAN GUNSTEREN, *A strategy for analysis of (molecular) equilibrium simulations: Configuration space density estimation, clustering, and visualization*, J. Chem. Phys., 114 (2001), pp. 2079–2089.

[21] W. L. JORGENSEN, J. CHANDRASEKHAR, J. D. MADURA, R. W. IMPEY, AND M. L. KLEIN, *Comparison of simple potential functions for simulating liquid water*, J. Chem. Phys., 79 (1983), pp. 926–935.

[22] M. E. KARPEN, D. J. TOBIAS, AND C. L. BROOKS, III, *Statistical clustering techniques for the analysis of long molecular dynamics trajectories: Analysis of* 2.2-ns *trajectories of YPGDV*, Biochemistry, 32 (1993), pp. 412–420.

[23] P. A. KOLLMAN, R. DIXON, W. CORNELL, T. VOX, C. CHIPOT, AND A. POHORILLE, *The development/application of a "minimalist" organic/biochemical molecular mechanic force field using a combination of* ab initio *calculations and experimental data*, in Computer Simulation of Biomolecular Systems, Vol. 3, A. Wilkinson, P. Weiner, and W. F. van Gunsteren, eds., Kluwer/Escom, The Netherlands, 1997, pp. 83–96.

[24] J. KUBELKA, J. HOFRICHTER, AND W. A. EATON, *The protein folding "speed limit,"* Curr. Opin. Struct. Biol., 14 (2004), pp. 76–88.

[25] S. KUMAR, D. BOUZIDA, R. H. SWENDSEN, P. A. KOLLMAN, AND J. M. ROSENBERG, *The weighted histogram analysis method for free-energy calculations on biomolecules.* I. *The method*, J. Comput. Chem., 13 (1992), pp. 1011–1021.

[26] R. E. KUNZ AND R. S. BERRY, *Statistical interpretation of topographies and dynamics of multidimensional potentials*, J. Chem. Phys., 103 (1995), pp. 1904–1912.

[27] P. LENZ, B. ZAGROVIC, J. SHAPIRO, AND V. S. PANDE, *Folding probabilities: A novel approach to folding transitions and the two-dimensional Ising-model*, J. Chem. Phys., 120 (2004), pp. 6769–6778.

[28] Y. LEVY, J. JORTNER, AND O. M. BECKER, *Dynamics of hierarchical folding on energy landscapes of hexapeptides*, J. Chem. Phys., 115 (2001), pp. 10533–10547.

[29] A. MA AND A. R. DINNER, *Automatic method for identifying reaction coordinates in complex systems*, J. Phys. Chem. B, 109 (2005), pp. 6769–6779.

[30] N. J. MARIANAYAGAM, N. L. FAWZI, AND T. HEAD-GORDON, *Protein folding by distributed computing and the denatured state ensemble*, Proc. Natl. Acad. Sci. USA, 102 (2005), pp. 16684–16689.

[31] D. MORONI, T. S. VAN ERP, AND P. G. BOLHUIS, *Investigating rare events by transition interface sampling*, Phys. A, 340 (2004), pp. 395–401.

[32] P. N. MORTENSON, D. A. EVANS, AND D. J. WALES, *Energy landscapes of model polyalanines*, J. Chem. Phys., 117 (2002), pp. 1363–1376.

[33] P. N. MORTENSON AND D. J. WALES, *Energy landscapes, global optimization and dynamics of the polyalanine Ac(ala)$_8$NHMe*, J. Chem. Phys., 114 (2001), pp. 6443–6454.

[34] S. B. OZKAN, K. A. DILL, AND I. BAHAR, *Fast-folding protein kinetics, hidden intermediates, and the sequential stabilization model*, Protein Science, 11 (2002), pp. 1958–1970.

[35] E. PACI, A. CAVALLI, M. VENDRUSCOLO, AND A. CAFLISCH, *Analysis of the distributed computing approach applied to the folding of a small β peptide*, Proc. Natl. Acad. Sci. USA, 100 (2003), pp. 8217–8222.

[36] V. S. PANDE, I. BAKER, J. CHAPMAN, S. P. ELMER, S. KHALIQ, S. M. LARSON, Y. M. RHEE, M. R. SHIRTS, C. D. SNOW, E. J. SORIN, AND B. ZAGROVIC, *Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing*, Biopolymers, 68 (2003), pp. 91–109.

[37] S. PARK AND V. S. PANDE, *Validation of Markov state models using Shannon's entropy*, J. Chem. Phys., 124 (2006), 054118.

[38] I. K. ROTERMAN, M. H. LAMBERT, K. D. GIBSON, AND H. A. SCHERAGA, *A comparison of the CHARMM, AMBER and ECEPP potentials for peptides.* II. $\phi - \psi$ *maps for N-acetyl alanine N'-methyl amide: Comparisons, contrasts and simple experimental tests*, J. Biomol. Struct. Dyn., 78 (1989), pp. 421–453.

[39] J. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, *Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes*, J. Comput. Phys., 23 (1977), pp. 327–341.

[40] V. Schultheis, T. Hirschberger, H. Carstens, and P. Tavan, *Extracting Markov models of peptide conformational dynamics from simulation data*, J. Chem. Theor. Comput., 1 (2005), pp. 515–526.

[41] Ch. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, *A direct approach to conformational dynamics based on hybrid Monte Carlo*, J. Comput. Phys., 151 (1999), pp. 146–168.

[42] Ch. Schütte and W. Huisinga, *Biomolecular conformations can be identified as metastable states of molecular dynamics*, in Handbook of Numerical Analysis—Special Volume on Computational Chemistry, Vol. X, P. G. Ciarlet and J.-L. Lions, eds., North–Holland, Amsterdam, 2002, pp. 699–744.

[43] D. Shalloway, *Macrostates of classical stochastic systems*, J. Chem. Phys., 105 (1996), pp. 9986–10007.

[44] M.-y. Shen and K. F. Freed, *Long time dynamics of met-enkephalin: Tests of mode-coupling theory and implicit solvent models*, J. Chem. Phys., 118 (2003), pp. 5143–5156.

[45] N. Singhal, J. D. Chodera, J. W. Pitera, V. S. Pande, K. A. Dill, and W. C. Swope, *An Automatic State Decomposition Method for the Construction of Discrete-State Markov Models of Protein Dynamics*, manuscript, 2006.

[46] N. Singhal, C. D. Snow, and V. S. Pande, *Using path sampling to build better Markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin*, J. Chem. Phys., 121 (2004), pp. 415–425.

[47] S. Sriraman, I. G. Kevrekidis, and G. Hummer, *Coarse master equation from Bayesian analysis of replica molecular dynamics simulations*, J. Phys. Chem. B, 109 (2005), pp. 6479–6484.

[48] Y. Sugita and Y. Okamoto, *Replica-exchange molecular dynamics method for protein folding*, Chem. Phys. Lett., 314 (1999), pp. 141–151.

[49] W. C. Swope, J. W. Pitera, and F. Suits, *Describing protein folding kinetics by molecular dynamics simulations: 1. Theory*, J. Phys. Chem. B, 108 (2004), pp. 6571–6581.

[50] W. C. Swope, J. W. Pitera, F. Suits, M. Pitman, M. Eleftheriou, B. G. Fitch, R. S. Germain, A. Rayshubski, T. J. C. Ward, Y. Zhestkov, and R. Zhou, *Describing protein folding kinetics by molecular dynamics simulations: 2. Example applications to alanine dipeptide and a beta-hairpin peptide*, J. Phys. Chem. B, 108 (2004), pp. 6582–6594.

[51] A. Ulitsky and D. Shalloway, *Variational calculation of macrostate transition rates*, J. Chem. Phys., 109 (1998), pp. 1670–1686.

[52] T. S. van Erp, D. Moroni, and P. G. Bolhuis, *A novel path sampling method for the calculation of rate constants*, J. Chem. Phys., 118 (2003), pp. 7762–7774.