Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics

John D. Chodera
Graduate Group in Biophysics, University of California, San Francisco, California 94143

Nina Singhal
Department of Computer Science, Stanford University, Stanford, California 94305

Vijay S. Pande
Department of Chemistry, Stanford University, Stanford, California 94305

Ken A. Dill
Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94143

William C. Swope[a]
IBM Almaden Research Center, 650 Harry Road, San Jose, California 95120

(Received 21 November 2006; accepted 13 February 2007; published online 19 April 2007)

To meet the challenge of modeling the conformational dynamics of biological macromolecules over long time scales, much recent effort has been devoted to constructing stochastic kinetic models, often in the form of discrete-state Markov models, from short molecular dynamics simulations. To construct useful models that faithfully represent dynamics at the time scales of interest, it is necessary to decompose configuration space into a set of kinetically metastable states. Previous attempts to define these states have relied upon either prior knowledge of the slow degrees of freedom or on the application of conformational clustering techniques which assume that conformationally distinct clusters are also kinetically distinct. Here, we present a first version of an *automatic* algorithm for the discovery of kinetically metastable states that is generally applicable to solvated macromolecules. Given molecular dynamics trajectories initiated from a well-defined starting distribution, the algorithm discovers long lived, kinetically metastable states through successive iterations of partitioning and aggregating conformation space into kinetically related regions. The authors apply this method to three peptides in explicit solvent—terminally blocked alanine, the 21-residue helical F_s peptide, and the engineered 12-residue β-hairpin trpzip2—to assess its ability to generate physically meaningful states and faithful kinetic models. © *2007 American Institute of Physics*. [DOI: 10.1063/1.2714538]

I. INTRODUCTION

Many biomolecular processes are fundamentally dynamic in nature. Protein folding, for example, involves the ordering of a polypeptide chain into a particular topology over the course of microseconds to seconds, a process which can go awry and can lead to misfolding or aggregation, causing disease.[1] Enzymatic catalysis may involve transitions between multiple conformational substates, only some of which may allow substrate access or catalysis.[2–4] Post-translational modification events, ligand binding, or catalytic events may alter the transition kinetics among multiple conformational states by modulating catalytic function, allowing work to be performed, or transducing a signal through allosteric change.[5–7] A purely static description of these processes is insufficient for mechanistic understanding—the dynamical nature of these events must be accounted for as well.

Unfortunately, these processes may involve molecular time scales of microseconds or longer, placing them well outside the range of typical detailed atomistic simulations employing explicit models of solvent. However, due to the presence of many energetic barriers on the order of the thermal energy, the uncertainty in initial microscopic conditions, and the stochasticity introduced into the system by the surrounding solvent in contact with a heat bath, any suitable description of conformational dynamics must *by necessity* be statistical in nature. This has motivated the development of stochastic kinetic models of macromolecular dynamics which might conceivably be constructed from short dynamics simulations, yet provide a useful and accurate statistical description of dynamical evolution over long times.

Several approaches have been used to construct these models. *Transition interface sampling*,[8] *milestoning*,[9] and methods based on commitment probability distributions[10,11] describe dynamics on a one-dimensional reaction coordinate, but can only be applied if an appropriate reaction coordinate can be identified such that relaxation transverse to this coordinate is fast compared to diffusion along it. Discrete-state, continuous-time master equation models, characterized by a matrix of phenomenological rate constants describing the rate of interconversion between states,[12] can be constructed by identifying local potential energy minima as states and

[a] Author to whom correspondence should be addressed. Electronic mail: swope@us.ibm.com

estimating interstate transition rates by transition state theory.[13–19] Unfortunately, the number of minima, and hence the number of states, grows exponentially with system size, making the procedure prohibitively expensive for larger proteins or systems containing explicit solvent molecules. Others have suggested that stochastic models of dynamics can be constructed by expansion of the appropriate dynamical operator in a basis set,[20–22] but this approach appears to be limited by the great difficulty of choosing rapidly convergent basis sets for large molecules, a process that is not fundamentally different from identifying the slow degrees of freedom.

Instead, much work has focused on the construction of discrete- or continuous-time Markov models to describe dynamics among a small number of states which may each contain many minima within large regions of configuration space.[23–33] In these models, it is hoped that a separation of time scales between fast *intrastate* motion and slow *interstate* motion allows the statistical dynamics to be modeled by stochastic transitions among the discrete set of metastable conformational states governed by first-order kinetics. Consider, for example, the isomerization of butane, which has three main metastable conformational states (*gauche-plus*, *gauche-minus*, and *trans*). At sufficiently low temperature, dynamics is dominated by long dwell times *within* each of these three states, punctuated by infrequent transitions between them. The slow interstate transition process is well described by first-order reaction kinetics for observation intervals longer than the fast molecular relaxation time for intrastate dynamics due to the presence of a separation of time scales.[34] Such a separation of time scales would be a natural consequence of the widely held belief that the nature of the energy landscape of biomacromolecules is hierarchical.[16,35–38] If the system reaches local equilibrium within the state before attempting to exit, the probability of transitioning to any other state will be independent of all but the current state. This allows the process to be modeled with either a discrete-time Markov chain (e.g., Ref. 26) or a continuous-time master equation model with coarse-grained time (e.g., Ref. 29). In either model, processes occurring on time scales faster than the time to reach equilibrium within each state cannot be resolved.

Markov models embody a concise description of the various kinetic pathways and their relative likelihood, facilitating comparison with experimental data and providing a powerful tool for mechanistic insight. Once the model is constructed and the time scale for Markovian behavior determined, it can be used to compute the stochastic temporal evolution of either a single macromolecule or a population of noninteracting macromolecules, allowing direct comparison of simulated and experimental observables for both single-molecule or ensemble kinetics experiments. In addition, useful properties difficult to access experimentally, such as state lifetimes,[39] relaxation from experimentally inaccessible prepared states,[40] mean first-passage times,[26] the existence of hidden intermediates,[41] and P_{fold} values or transmission coefficients,[42] can easily be obtained. This allows for both a thorough understanding of mechanism and the generation of new, experimentally testable hypotheses.

To build such a model, it is necessary to decompose configuration space into an appropriate set of metastable states. If the low-dimensional manifold containing all the slow degrees of freedom is known *a priori*, then this can be partitioned into free energy basins to define the states, such as by examination of the potential of mean force.[25,28,29,32,40] In the absence of this knowledge, others have turned to conformational clustering techniques to identify conformationally distinct regions which may also be kinetically distinct.[24,26,27,43]

Instead, we adopt a strategy first suggested for the discovery of metastable states in biomolecular systems by researchers at the Konrad-Zuse-Zentrum für Informationstechnik.[44] The principal idea is this: If configuration space could be decomposed into a large number of small cells, the probability of transitioning between these cells in a fixed evolution time could be measured. This probability is a measure of *kinetic connectivity* among the cells, which allows the identification of aggregates of these cells that approximate true metastable states.[45] Unfortunately, the choice of how to divide configuration space into cells is not straightforward. Suppose one is to consider the analysis of some fixed amount of simulation data. If configuration space is decomposed very finely, the boundaries between metastable states can in principle be well approximated, but the estimated cell-to-cell transition probabilities will become statistically unreliable. On the other hand, if configuration space is decomposed too coarsely, the transition probabilities may be well determined, but the boundaries between metastable states cannot be resolved clearly, potentially disrupting or destroying the Markovian behavior of interstate dynamics. An optimal choice would ultimately require knowledge of the metastable regions in order to determine the best decomposition of configuration space into cells.

In this work, we propose an iterative procedure to determine both the choice of cells and their aggregates to approximate the desired metastable states. We use a conformational clustering method to carve configuration space into an initial crude set of cells (*splitting*) and a Monte Carlo simulated annealing procedure to collect metastable collections of cells into states (*lumping*). This cycle is repeated, with the splitting procedure now applied individually to each state to generate a new set of cells, and the lumping procedure applied to the entire set of cells to redefine states until further application of this procedure leaves the approximations to metastable states unchanged. This procedure allows state boundaries to be iteratively refined, as regions that mistakenly have been included in one state can be split off and regrouped with the proper state. Throughout this process, we require that the cells never become so small that estimation of the relevant transition matrix elements is statistically unreliable. Our proposed method is efficient, of $\mathcal{O}(N)$ complexity in the number of stored configurations, and can easily be parallelized.

This paper is organized as follows: In Sec. II, we give an overview of the Markov chain model and its construction, elaborate on desirable properties of an algorithm to partition configuration space into states, and outline the principles underlying the algorithm we present here. In Sec. III, we provide a detailed description of the automatic state decomposition algorithm and its implementation. In Sec. IV, we apply this algorithm to three model peptide systems in explicit solvent to assess its performance: alanine dipeptide, the 21-residue F_s helix-forming peptide, and the 12-residue engineered trpzip2 hairpin. Finally, in Sec. V, we discuss the advantages and shortcomings of our algorithm, with the hope that future state decomposition algorithms can address the remaining challenges.

II. THEORY

Some discussion of the stochastic model of kinetics considered here and the theory underlying the method is appropriate before describing the algorithmic implementation in detail. The actual implementation of the algorithm used here is described in detail in Sec. III.

A. Markov chain and master equation models of conformational dynamics

Consider the dynamics of a macromolecule immersed in solvent, where the solvent is at equilibrium at some particular temperature of interest. We presume that all of configuration space has already been decomposed into a set of nonoverlapping regions, or *states*, which together form a complete decomposition of configuration space. The method by which these states are identified is described in subsequent sections.

If we observe the evolution of this system at times $t=0,\tau,2\tau,\ldots$, where τ denotes the observation interval, we can represent this sequence of observations in terms of the state the system visits at each of these discrete times. The sequence of states produced is a realization of a *discrete-time stochastic process*. For this process to be described by a Markov chain, it must satisfy the *Markov property*, whereby the probability of observing the system in any state in the sequence is independent of all but the previous state. For a stationary process on a finite set of L states, this process can be completely characterized by an $L\times L$ *transition matrix* $\mathbf{T}(\tau)$ dependent only on the observation interval or *lag time* τ. (We adopt the notation for a *column-stochastic* transition matrix, in which the columns sum to unity; this differs from the notation in some previously cited references, which use a *row-stochastic* transition matrix, equal to the transpose of the column-stochastic matrix used here.) The element $T_{ji}(\tau)$ denotes the probability of observing the system in state j at time t given that it was previously in state i at time $t-\tau$. If this process satisfies detailed balance (which we will assume to be the case for physical systems of the sort we consider here[12]) we additionally have the requirement

$$T_{ji}p_{\mathrm{eq},i} = T_{ij}p_{\mathrm{eq},j}, \tag{1}$$

where $p_{\mathrm{eq},i}$ denotes the equilibrium probability of state i.

The vector of probabilities of occupying any of the L states at time t (here also referred to as the vector of state populations, such as in an experiment involving a population of noninteracting macromolecules) can be written as $\mathbf{p}(t)$. If the initial probability vector is given by $\mathbf{p}(0)$, we can write the probability vector at some later time $n\tau$ as

$$\mathbf{p}(n\tau) = \mathbf{T}(n\tau)\mathbf{p}(0) = [\mathbf{T}(\tau)]^n\mathbf{p}(0). \tag{2}$$

This is a form of the *Chapman-Kolmogorov equation*.

Alternatively, the process can be characterized in *continuous* time by a matrix of phenomenological rate constants $\mathbf{K}$, where the element K_{ji}, $j\neq i$ denotes the non-negative phenomenological rate from state i to state j. The diagonal elements are determined by $K_{ii}=-\Sigma_{j\neq i}K_{ji}$ to ensure the columns sum to zero so as to conserve probability mass. Time evolution is then governed by the equation

$$\dot{\mathbf{p}}(t) = \mathbf{K}\mathbf{p}(t), \tag{3}$$

where the dot represents differentiation with respect to time. This evolution equation has the formal solution

$$\mathbf{p}(t) = e^{\mathbf{K}t}\mathbf{p}(0), \tag{4}$$

where the exponential denotes the formal matrix exponential. Equation (3) is often referred to as a *master equation*[12,46] describing evolution among a discrete set of states in continuous time. It is important to note that, despite the fact that $\mathbf{p}(t)$ is formally defined for all times t, we do not expect Eq. (4) to hold for *all* times t for physical systems of the sort we consider here. In particular, for states of finite extent in configuration space, there exists a corresponding limit for the time resolution for which dynamics will appear Markovian; processes that occur on time scales shorter than this will be incorrectly described by the master equation.

There is an obvious relationship between the transition matrix $\mathbf{T}(\tau)$ and the rate matrix $\mathbf{K}$ evident from comparison of Eqs. (2) and (4),

$$\mathbf{T}(\tau) = e^{\mathbf{K}\tau}. \tag{5}$$

If the process can be described by a continuous-time Markov process at all times, then this process can be equivalently described at discrete time intervals by the corresponding transition matrix. The converse may not always be true due to sampling errors in $\mathbf{T}(\tau)$, though methods exist to recover rate matrices $\mathbf{K}$ consistent with the observed data and the requirements of detailed balance and nonnegative rates.[23,29]

The transition and rate matrices have eigenvalues $\mu_k(\tau)$ and λ_k, respectively, and share corresponding right eigenvectors $\mathbf{u}_k$. The detailed balance requirement additionally ensures that all eigenvalues are real, and we here presume them to be sorted in descending order. $\mu_k(\tau)$ and λ_k are related by

$$\mu_k(\tau) = e^{\lambda_k\tau}. \tag{6}$$

The eigenvalues each imply a time scale

$$\tau_k = -\lambda_k^{-1} = -\tau[\ln \mu_k(\tau)]^{-1}, \tag{7}$$

and the associated eigenvector gives information about the aggregate conformational transitions that are associated with this time scale.[44,45,47,48] In particular, the components of $\mathbf{u}_k$ sum to zero for each $k\geqslant 2$, and the aggregate dynamical mode corresponds to transitions from states with positive eigenvector components to states with negative components and vice versa, with the degree of participation in the mode governed by the magnitude of the eigenvector component. This property can be useful in identifying metastable states.

For the remainder of this manuscript, we will refer exclusively to the discrete-time Markov chain model picture without loss of generality [Eq. (2)].

B. Markov model construction from simulation data given a state partitioning

Once a statistical-mechanical ensemble describing equilibrium and a microscopic model describing dynamical evolution in phase space have been selected, the transition matrix $\mathbf{T}(\tau)$ can be estimated from molecular dynamics simulations. For a system in which dynamical evolution is Newtonian and, at equilibrium, configurations are distributed according to a canonical distribution at a given temperature, Swope *et al.*[39] show that the transition probability $T_{ji}(\tau)$ can be written as the following ratio of canonical ensemble averages:

$$T_{ji}(\tau) = \frac{\int d\mathbf{z}(0) e^{-\beta H(\mathbf{z}(0))} \chi_j(\mathbf{z}(\tau)) \chi_i(\mathbf{z}(0))}{\int d\mathbf{z}(0) e^{-\beta H(\mathbf{z}(0))} \chi_i(\mathbf{z}(0))} \tag{8}$$

$$= \frac{\langle \chi_j(\tau) \chi_i(0) \rangle}{\langle \chi_i \rangle}, \tag{9}$$

where $\mathbf{z}(t)$ denotes a point in phase space visited by a trajectory at time t, $\chi_i(\mathbf{z})$ denotes the indicator function for state i (which assumes a value of unity if $\mathbf{z}$ is in state i, and zero otherwise), $\beta \equiv (k_B T)^{-1}$ the inverse temperature, $H(\mathbf{z})$ the Hamiltonian, and $\langle A \rangle$ the canonical ensemble expectation of a phase function $A(\mathbf{z})$ at inverse temperature β.

Given a set of simulations initiated from an equilibrium distribution, the expectations in Eq. (9) can be computed independently by standard analysis methods.[49] Estimation of the correlation function in the numerator can make use of both the stationarity of an equilibrium distribution (by considering overlapping intervals of time τ) and the microscopic reversibility (by considering also time-reversed versions of the simulations) of Newtonian trajectories. Alternatively, if an equilibrium distribution within each state can be prepared, one can also directly estimate a column of transition matrix elements by computing the fraction of trajectories initially at equilibrium within state i that terminate in state j a time τ later. More elaborate methods based on equilibrium ensembles prepared within special *selection cells* that are not coincident with the states[25,39] or *partition of unity* restraints[50] can also be used to compute transition matrix elements efficiently.

C. Requirements for a useful Markov model

For any given state partitioning, the dynamics of the system will be Markovian on some time scale. For example, if the lag time τ is so long as to approach the time for the system to relax to an equilibrium distribution from any arbitrary starting distribution, a single application of the transition matrix $\mathbf{T}(\tau)$ produces the invariant equilibrium distribution. However, if this τ exceeds the time scale of the process of interest, our model is not useful for describing it, and therefore it is advantageous to attempt to find a state decomposition that is Markovian on a shorter time scale in order to extract useful dynamical information about this process. (Equilibrium probabilities can still be extracted from the stationary eigenvector, the eigenvector corresponding to an eigenvalue of unity, of such a transition matrix, which may have some utility if one had constructed the transition matrix from trajectories not initiated from distributions at equilibrium globally.)

For a given state i, we will define its internal equilibration time, $\tau_{\mathrm{int},i}$, as the characteristic time one must wait before the system, initially in a configuration within state i, generates a new *uncorrelated* configuration within the state by dynamical evolution. This internal equilibration time, or *memory time*, closely related to the molecular relaxation time scale τ_{mol} in Chandler's reactive flux formulation of transition state theory,[34] depends, of course, on the choice of state decomposition. We can denote the longest of these times over all states by τ_{int}. If the lag time is longer than τ_{int}, we will expect the system to have lost memory of its previous location within *any* state it may have been in, either remaining within that state or transitioning to a new one, and for dynamics on this set of states to be independent of history. On the other hand, for lag times shorter than τ_{int}, we cannot guarantee that transition probabilities are independent of history everywhere. This suggests a way in which the utility of various decompositions can be measured. For a fixed number of states, the most useful model will partition configuration space to yield the shortest τ_{int}, as this model can be used to study the widest range of dynamical processes.

In addition to producing transition probabilities that are history independent at a relevant lag time, we impose an additional condition on our states to ensure the resulting model also provides physical and chemical insight. In order for the states to be defined such that equilibration within a state is rapid, we desire that the region of configuration space defining each state be *connected*. A state composed of two or more unconnected regions of configuration space defies the assumption that equilibration within the state is much faster than the characteristic time to leave it.

D. Validation of Markov models

Once a decomposition of configuration space is chosen, we are faced with the task of determining the observation time interval τ at which dynamics in this state space appears Markovian. Unfortunately, we cannot directly compute the internal state equilibration times, though examination of the eigenvalues of the transition matrix restricted to a state may give a lower bound on this time in the absence of statistical uncertainty.[51] The most rigorous test for Markovian behavior would be a direct check of history independence. The simplest test of this type is to compute second-order transition probabilities and compare them to the appropriate products of the first-order transition probabilities to see if their disagreement is statistically significant. While it is possible to estimate the second-order probabilities from the simulation data, this requires the estimation of three-time correlation functions, which often possess statistical uncertainties so large as to render them useless for this kind of test.[52] Additionally, this would miss possible yet unlikely higher order history dependencies.

1. Information-theoretic metric

Another approach, from Park and Pande[33] uses concepts from information theory to compute the *conditional mutual information* conveyed by the second-to-last state, which quantifies the discrepancy between observed second-order transition probabilities and the estimate modeled from first-order transition probabilities. The result of this analysis is a scalar that quantifies the degree of history dependence. For a pure first-order Markov process, the mutual information will be zero, as no additional information is gained by including additional history. While this method also requires computing three-time correlation functions, which may individually have substantial uncertainties, the weighted combination of these into a single value reduces the uncertainty in the resulting metric. Unfortunately, there is no rigorous criteria for how small this measure must be in order for the model to be considered acceptably Markovian.

2. Chapman-Kolmogorov

Alternatively, raising the transition matrix to a power n (hence summing over the intermediate states) and comparing with the observed transition probabilities for a lag time of $n\tau$, such that one is effectively determining whether the Chapman-Kolmogorov equation [Eq. (2)] is satisfied, helps to reduce the uncertainty so that the test becomes practical. This is equivalent to propagating the population in time out of a probability distribution confined to each state i initially and comparing the model evolution with the observed transition probabilities over times much longer than τ_{int}. This serves as a check to ensure that the model is at least consistent with the data set from which it was constructed, to within the statistical uncertainty of the transition matrices obtained from the data set. This method was employed, for example, in Refs. 39 and 40 and is used here as well.

3. Implied time scales

Swope *et al.*[39] suggested a number of additional tests for signatures of Markov behavior, the most sensitive of which appears to be examining the behavior of the *implied time scales* of the transition matrix $\mathbf{T}(\tau)$, which can be computed from the eigenvalues of the transition matrix by Eq. (7), as a function of increasing lag time τ.[52] At sufficiently large τ, the implied time scales will be independent of τ, implying that exponentiation of the transition matrix is nearly identical to constructing the transition matrix using longer observation time intervals [Eq. (2)]. The shortest observation time interval for which this holds can be correlated with the internal equilibration time τ_{int}, and descriptions of the behavior of the system using that state decomposition should be Markovian for all lag times $\tau \geqslant \tau_{\text{int}}$. This is also a test of whether the Chapman-Kolmogorov equation holds, but as it computes only L numbers and orders them by time scale, it allows emphasis to be placed on the longest time scales in the system. Implied time scales were used for all systems considered here.

Unfortunately, this last method has some drawbacks. First, small uncertainties in the eigenvalues of the transition matrix can induce very large uncertainties in the implied time scales. With increasing lag time τ, the number of statistically independent observed transitions from which $\mathbf{T}(\tau)$ is estimated diminishes, and the statistical uncertainty in the implied time scales τ_k will grow. Second, while stability of the implied time scales with respect to lag time is a *necessary* consequence of history independence, it is not itself *sufficient* to guarantee history independence, though we may be unlikely to encounter physical systems for which this is problematic. However, tests on simple models indicate that the information-theoretic metric suggests the emergence of Markovian behavior on similar lag times to this method, suggesting some degree of fundamental equivalence.[33]

III. THE AUTOMATIC STATE DECOMPOSITION ALGORITHM

Based on the theory above, we provide a list of practical considerations for an automatic state decomposition algorithm and then present an algorithm that meets them. The algorithm operates on an ensemble of molecular dynamics trajectories where conformations have been stored at regular time intervals. In this work, we apply the method to a set of *equilibrium* trajectories at the temperature of interest, but the algorithm can in principle be applied to trajectories generated from *biased* initial conditions, provided the unbiased transition probabilities between regions of configuration space can be computed. We stress that the algorithm presented here is simply a first attempt at a truly general and automatic algorithm for use with biomacromolecules.

A. Practical considerations for an automatic state decomposition algorithm

There are several desirable properties that a state decomposition should possess to be both useful and practical.

(1) It is not uncommon for simulations conducted on supercomputers such as Blue Gene,[53,54] distributed computing platforms such as Folding@Home,[55,56] or even computer clusters to generate data sets that may contain 10^5-10^7 configurations in up to 10^4 trajectories, therefore rendering impractical the use of any algorithm with a computational complexity greater than $\mathcal{O}(N \log N)$ in the number of configurations.

(2) We assume configurations lie exclusively in the configuration space of the macromolecule. We presume decorrelation of momenta and reorganization of the solvent is faster than the processes of interest. (We recognize that solvent coordinates may be critical in some phenomena, but dealing with solvent degrees of freedom would also require accounting for the indistinguishability of solvent molecules upon their exchange. We leave this to further versions of the algorithm.)

(3) Molecules may have symmetries due to the presence of chemically equivalent atoms such as in aromatic rings, methyl protons, and the oxygens of carboxylate groups. The state decomposition should be invariant to permutations of these atoms.

(4) The state decomposition algorithm should produce a decomposition for which dynamics appears to be Markovian at the shortest possible lag time τ, so as to produce the most useful model.

(5) The resulting model should not include so many states so that the elements of the transition matrix will be statistically unreliable.

B. Sketch of the method

A state decomposition algorithm intended to produce the most *useful* Markov models, as discussed in Sec. II C above, would generate models that minimize the internal equilibration time τ_{int}, the minimum time for which the model behaves in a Markovian fashion. If states can be constructed where the time scale for equilibration *within* each state is much shorter than the time scale for transitions *among* the states, we would expect interstate dynamics to be well modeled by a Markov chain after sufficiently long observation intervals. Unfortunately, τ_{int} is difficult to determine directly, so we are instead forced to identify some surrogate quantity whose maximization will hopefully lead to improved separation between the time scales for intrastate and interstate transitions. Following the approach of Ref. 57, we define a measure of the *metastability* Q of a partitioning into L *macrostates* as the sum of the self-transition probabilities for a given lag time τ,

$$Q \equiv \sum_{i=1}^{L} T_{ii}(\tau). \tag{10}$$

For $\tau=0$, $Q=L$, and Q decays to unity as τ grows large enough for the self-transition probabilities T_{ii} to reach the equilibrium probabilities of each macrostate. Poor partitionings will result in a small Q, as trajectories started in some states will rapidly exit; conversely, good partitionings into strongly metastable states will result in a large Q, as trajectories will remain in each macrostate for long times. In the absence of statistical uncertainty, Q is bounded from above by the sum of the L largest eigenvalues of the true dynamical propagator for the system.[57]

The goal of our algorithm is to identify a partitioning into L contiguous macrostates that maximizes the metastability Q. While in principle, the boundaries between these macrostates can be varied directly to optimize Q, in analogy to variational transition state theory,[58] a complicated parameterization may be necessary to describe the potentially highly convoluted hypersurfaces separating the states, and Q may have multiple maxima in these parameters. Instead, we choose an approach based on *splitting* the conformation space into a large number of small contiguous *microstates* and then *lumping* these microstates into macrostates to maximize the metastability.

This approach is similar to the approach of Schütte *et al.* described in Ref. 44, but with a substantial difference. In their work, each degree of freedom of the molecule (such as a torsion angle) is subdivided independently to produce a multidimensional grid. As the number of states is exponential in the number of degrees of freedom, this approach quickly becomes intractable for macromolecules that possess large numbers of degrees of freedom, even if the sparsity of the transition matrix is taken into account. Instead, we choose to let the data define the low-dimensional manifold of configuration space accessible to the macromolecule, and we can apply any clustering algorithm that is $\mathcal{O}(N \log N)$ in the number of configurations to decompose the sampled conformation space into a set of K contiguous microstates. This step corresponds to the first *split* step in Fig. 1.

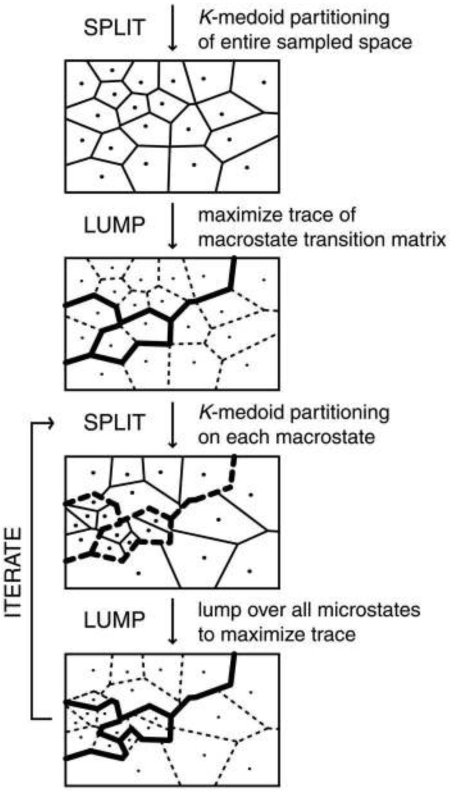
SPLIT
K-medoid partitioning of entire sampled space
LUMP
maximize trace of macrostate transition matrix
SPLIT
K-medoid partitioning on each macrostate
ITERATE
LUMP
lump over all microstates to maximize trace

FIG. 1. Flow chart of the automatic state decomposition algorithm. We consider K microstates which are used as the basis to construct $L<K$ macrostates that are the approximations to the true metastable states in the system.

Once the conformation space is divided into K microstates, we *lump* the microstates together to produce $L<K$ macrostates with high metastability Q. This corresponds to the first *lump* step in Fig. 1. The difficulty here is that the uncertainty in the metastability of a partitioning can be large if any macrostate contains very few configurations. Since a macrostate may consist of a single microstate, the microstates must be large enough for the self-transition elements to be statistically well determined. This comes at a price: with large microstates, the procedure may have difficulty accurately determining the boundaries between macrostates because the resolution of partitioning is limited by the finite extent of the microstates. Additionally, the choice of decomposition into microstates is arbitrary, whereas we would like the state decomposition algorithm to produce equivalent sets of macrostates regardless of the quality of the initial partitioning.

To overcome these difficulties, we *iterate* the aforementioned procedure. After microstates are combined into macrostates, each macrostate is again fragmented into a new set of microstates (the second *split* step in Fig. 1). The refined set of all microstates is then lumped to form refined macrostates (the second *lump* step in Fig. 1). In this way, the boundaries between macrostates are iteratively refined, and regions incorrectly lumped in previous iterations may be split off and lumped with the correct macrostate in subsequent iterations. At convergence, no shuffling of conformations between macrostates should occur.

There is unfortunately no unambiguous way to choose the number of states L. If there is a clean separation of time scales, examination of the eigenvalue spectrum of the microstate transition matrix may suggest an appropriate value of L.[45] In a hierarchical system, there will be many gaps in the eigenvalue spectrum and many choices of L will lead to good Markovian models of varying complexity. There is, however, a tradeoff between the number of states and the amount of data needed to obtain a model with the same statistical precision. It may be necessary to apply the algorithm repeatedly with different choices of L to produce a model adequate for describing the time scales of interest. L could even be chosen dynamically at each iteration of the algorithm, though we did not choose to do so in this version.

C. Implementation

There are a number of implementation choices to be made in the algorithm given above, and here we briefly summarize and justify our selections.

1. Splitting

For the split step, we choose to apply K-medoid clustering[59] for a fixed number of iterations because of its $\mathcal{O}(KN)$ time complexity (where K can be taken to be constant) and ease of parallelization. Additionally, K-medoid clustering has an advantage over the more popular K-means clustering[60] in this application, as it does not require averaging over conformations, which may produce nonsensical constructs when drastically different conformations are included in the average. Splitting by K-medoid clustering is initiated from a random choice of K unique conformations to function as *generators*. All conformations are assigned to the microstate identified by the generator they are closest to by some distance metric (defined below). Next, an attempt is made to update the generator of each microstate. K members of the microstate, drawn at random, are evaluated to see if they reduce the intrastate variance of some distance metric from the generator. If so, the configuration for which the intrastate variance is minimal is assigned as the new generator. All conformations are then reassigned to the closest generator, and the process of updating the generators is repeated. In standard K-medoid applications, this procedure is iterated to convergence, but since the purpose of the splitting phase is simply to divide the sampled manifold of configuration space into contiguous states, ensuring that each state is significantly populated, only five iterations of this procedure were used.

For the distance metric, we selected the root-mean squared deviation (RMSD), computed after a minimizing rigid body translation and rotation using the rapid algorithm of Theobald.[61] In the first splitting iteration, only C_α atoms were used to compute the RMSD due to the expense of having to cluster all conformations in the data set; in subsequent iterations, all heavy atoms (excepting those indistinguishable by symmetry) were used, as well as side chain polar hydrogens. This metric was chosen because it possesses all the qualities of a proper distance metric,[62] accounts for both local similarities between pairs of conformations as well as global ones, and runs in time proportional to the number of atoms, as opposed to a metric such as distance matrix error (or dRMSD), which scales as the square of the number of atoms. In molecules with additional symmetry, the distance metric can be adjusted accordingly. Our choice of distance metric is not the only one that would suffice; any distance metric which can distinguish between kinetically distinct conformations is sufficient for this algorithm. In constrast, using something like backbone RMSD throughout the process may be a poor distance metric since it would ignore potentially relevant side chain kinetics.

2. Lumping

Lumping to L states so as to maximize the metastability Q of the macrostates proceeds in two stages. In the first stage, information on the metastable state structure contained in the eigenvectors associated with the slowest time scales[45,47,48,63] is used to construct an initial guess at the optimal lumping. Because the eigenvectors contain statistical noise, this may not actually be optimal, so we include a second stage that uses a Monte Carlo simulated annealing (MCSA) optimization algorithm to further improve the metastability. Though the MCSA algorithm could in principle be used without the first stage to find optimal lumpings, we find its convergence is greatly accelerated by use of the initial guess. Ensuring connectivity during the lumping stage would be difficult due to the need to enumerate neighbors in configuration space, but in practice, we find this unnecessary.

In the first stage, a transition matrix among microstates is computed [using Eq. (9)] taking advantage of both stationarity and time reversibility for a short lag time τ, typically the shortest interval at which configurations were stored. Motivated by the Perron cluster cluster analysis algorithm of Deuflhard *et al.*,[63] an initial guess for the optimal lumping of microstates to macrostates is generated using the *left* eigenvectors (the left eigenvector $\mathbf{v}_k$ is simply related to the right eigenvector $\mathbf{u}_k$ by $(\mathbf{v}_k)_i = p_{\mathrm{eq},i}^{-1}(\mathbf{u}_k)_i$[46]) associated with the largest eigenvalues of the microstate transition matrix. We begin by assigning all microstates to a single macrostate. For each eigenvalue, the corresponding eigenvector contains information about an aggregate transition between the set of microstates with positive eigenvector components and the set with negative components, with a time scale determined by the eigenvalue. Equilibration within each set must occur on a faster time scale, provided the eigenvalues are nondegenerate. We can therefore use this information to identify one macrostate to divide in two. We select the macrostate with the largest L_1 norm of eigenvector components (restricted to microstates belonging to the macrostate) after subtracting the mean of these components. In Ref. 63, the sign structure alone was used to split these sets, but since we restrict the splitting to a single macrostate, we split about the mean, so that microstates with eigenvector components above the mean become one macrostate and the rest go into another. This procedure is performed for eigenvectors $k=2,\ldots,L$ in order, which should correspond to the slowest processes in the system, generating a total of L macrostates.

Due to statistical noise in the eigenvectors and near degeneracy in the eigenvalues, this procedure does not always result in the lumping with the maximal metastability Q. Therefore, in the second stage, the metastability was maximized using a MCSA algorithm, using the eigenvector-generated lumping as an initial seed. In each step of the Monte Carlo procedure, a microstate was selected with uniform probability and assigned to a random macrostate. If this proposed move would leave a macrostate empty or did not change the partitioning, it was rejected immediately. The proposed partitioning was accepted with probability $\min\{1,e^{\beta\Delta Q}\}$. The effective inverse temperature parameter β was set to be equal to the step number, and the MCSA procedure run for 20 000 steps. Twenty independent MCSA runs were initiated from the initial eigenvector-based partitioning, and the partitioning with the highest metastability sampled in any run was selected to define the lumping into macrostates. No attempt was made to optimize the annealing schedule.

It should be noted that the metastability Q is not the only surrogate that could be optimized in order to produce a useful state decomposition. One could choose to maximize the largest eigenvalues or fastest time scales of the lumped transition matrix, the product of eigenvalues (which would give more weight to faster time scales), or even a weighted sum of the eigenvalues, where the weights might be due to the equilibrium importance of the eigenmode in dynamics or in modeling a process of interest. Unfortunately, these quantities all necessitate computing some eigenvalues or the determinant of the lumped transition matrix for every proposed lumping to be evaluated by the MCSA algorithm, which would add a significant computational burden. Alternatively, other quantities could be computed from the transition matrix directly, such as the state lifetimes estimated from the self-transition probabilities as $\tau_{L,i}=(1-T_{ii})^{-1}$. However, the combination of computational and theoretical convenience makes the use of metastability a natural choice here.

3. Iteration

For the remaining iterations, the K-medoid clustering is repeated independently on each macrostate for five iterations. In general, we split each macrostate into ten microstates, unless otherwise noted. However, we wish to ensure statistical reliability of the transition probability matrix. If the expected microstate size (estimated by the population of the macrostate divided by K) falls below some threshold (100 configurations unless otherwise noted), we split to a smaller number of states such that the expected size is above the threshold. The lumping step is then repeated on all resulting microstates. The entire procedure of splitting and lumping is repeated for a total of ten iterations, which for the applications considered here was sufficient for convergence of the metastability.

D. Validation

To validate the model, we examine the largest implied time scales as a function of lag time, as computed from the eigenvalues of the transition matrix by Eq. (7). In particular, we attempt to determine the minimum lag time after which the implied time scales appear to be independent of lag time to within the estimated statistical uncertainty (see Sec. II D). To estimate statistical uncertainties in the implied time scales and other quantities, we perform a bootstrap procedure[64] on the pool of independent trajectories. Forty bootstrap replicates, each consisting of a number of trajectories equal to the number of independent trajectories in the data set pool, are generated by drawing from the pool with replacement. For alanine dipeptide, 100 bootstrap replicates were used. For each replicate, the implied time scales or other quantity is computed, and either the standard deviation over the sample of replicates computed (if reported in the text as $a\pm b$) or a 68% confidence interval centered on the sample mean estimated (if depicted in a figure as vertical error bars).

We also estimate the number of statistically independent visits to each macrostate. Since sequential samples from a single trajectory are temporally correlated, we compute the integrated autocorrelation time[65,66] $\tau_{\mathrm{ac},i}$ for each macrostate i. Ignoring statistical uncertainty, this correlation time is an upper bound on the equilibration time within a state; long-lived states will necessarily have long autocorrelation times, but trajectories trapped within them may contain many uncorrelated samples if the internal equilibration time is short. In the absence of a convenient way to quantify the internal equilibration time for each state, the autocorrelation time provides a better estimate of the appropriate time scale than the time to reach global equilibrium τ_{eq}. The effective number of independent samples for each state is estimated by summing the number of independent samples from each trajectory (which are assumed independent), where the effective number of independent samples of state i from trajectory n is computed as $N_{ni}^{\mathrm{eff}}\approx\min\{1,N_{ni}/g_i\}$, where N_{ni} is the number of configurations from trajectory n in state i, and $g_i=1+2\tau_{\mathrm{ac},i}/\tau_{\mathrm{sample}}$ is the statistical inefficiency of state i, where τ_{sample} is the sampling interval between conformations.

IV. APPLICATIONS

A. Alanine dipeptide

We first demonstrate the application of the automatic state decomposition algorithm to a simple model system, terminally blocked alanine peptide (sequence Ace-Ala-Nme) in explicit solvent. Because the slow degrees of freedom (ϕ and ψ torsions, labeled in Fig. 2, left) are known *a priori* (simulations of alanine dipeptide examining the committor distribution have implicated solvent coordinates as the next-slowest degrees of freedom,[68,71] but we have previously verified that ϕ and ψ torsions form a sufficient basis for the slow degrees of freedom on time scales of 6 ps and greater[40]), it is relatively straightforward to manually identify metastable states from examination of the potential of mean force, making it a popular choice for the study of biomolecular dynamics.[17,40,67–70] Previously, a master equation model constructed using six manually identified states (Fig. 2, right)

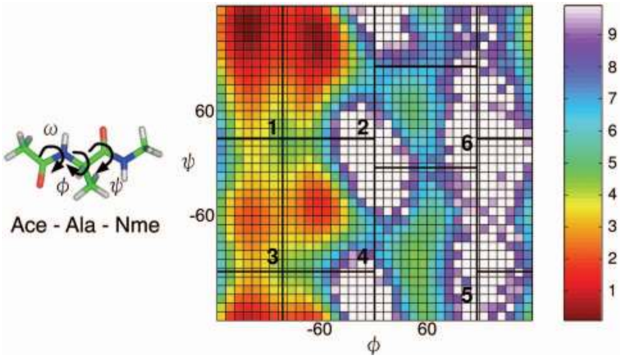

FIG. 2. (Color) Potential of mean force and manual state decomposition for alanine dipeptide. Left: The terminally blocked alanine peptide with ϕ, ψ, and ω backbone torsions labeled. Right: The potential of mean force in the (ϕ,ψ) torsions at 400 K estimated from the parallel tempering simulation, truncated at $10k_BT$ (white regions), with reference scale (far right) labeled in units of k_BT. Boundaries defining the six states manually identified in Ref. 40 from examining the 300 K PMF are superimposed, and the states labeled.

was shown to reproduce dynamics over long times (with the time to reach equilibrium over 100 ps at 302 K) given trajectories only 6 ps in length.[40] We therefore determine whether the automatic algorithm can recover a model of equivalent utility to this manually constructed six-state decomposition for this system, as well as study its convergence properties. Because the algorithm uses the solute Cartesian coordinates, rather than the (ϕ,ψ) torsions, this is a good test of whether good approximations to the true metastable states can be discovered without prior knowledge of the slow degrees of freedom. For ease of visualization, however, we project the state assignments onto the (ϕ,ψ) torsion map for comparison with our manually constructed states.

1. Simulation details

Trajectories were obtained from the 400 K replica of a 20 ns/replica parallel tempering simulation (note that only 10 ns/replica were used in Ref. 40—the data presented here includes an additional 10 ns/replica of production simulation; additionally configurations containing *cis*-ω torsions discussed in the text were not observed in the first 10 ns/replica cited in the previous study—these conformations only appeared in the latter 10 ns/replica) described in Ref. 40, and consisted of an equilibrium pool of 1000 constant-energy, constant-volume trajectory segments 20 ps in length with configurations stored every 0.1 ps. The peptide was modeled by the AMBER 96 forcefield[72] and solvated in TIP3P water.[73] The previous study[40] considered the dynamics at 302 K, but resorted to a focused sampling strategy where a number of trajectories were initiated from equilibrium distributions within constricted *selection cells*[39] in order to obtain statistically reliable estimates of the transition matrix. Here, as the focus was on locating these metastable states from equilibrium data, we found it necessary to use equilibrium data from a higher temperature—here, the 400 K replica—in order to obtain sufficient numbers of trajectories covering the entirety of the landscape. A two-dimensional potential of mean force (PMF) at 400 K in the (ϕ,ψ) backbone torsions was estimated from the parallel tempering simulation using the weighted histogram analysis method[74,75] by discretizing each degree of freedom into 10° bins (Fig. 2). Because the (ϕ,ψ) torsions are supposed to be the *only* slow degrees of freedom in the system, we can associate basins in the potential of mean force with metastable states. The six such states identified from the 302 K PMF in the previous study,[40] identified as dark lines in Fig. 2, can be seen to adequately separate the free energy basins observed at 400 K. We take this decomposition as our reference "gold standard" and compare the one obtained from our automatic state decomposition algorithm with it.

2. Automatic state decomposition

First, the automatic state decomposition method described in Sec. III was applied to this dataset in a fully automatic way to obtain six macrostates that could be compared with the gold standard. Since there is only one C_α atom in the peptide, we opted to use the backbone RMSD (including the amide proton and carbonyl oxygen) in the first stage, splitting to 100 microstates; subsequent iterations used the distance metric and splitting procedure described in Sec. III C. A single sampling interval—0.1 ps—was used for the calculation of the metastability metric Q used in lumping, as described in Sec. III B. Application of the state decomposition algorithm to the entire data set revealed a state that heavily overlapped with several others when projected onto the (ϕ,ψ) map, along with an extremely long time scale associated with its transitions (data not shown). Closer examination of the ensembles of configurations contained in this overlapping state revealed that the overlapping regions differed by a peptide bond isomerization; a small population of the trajectories contained a N-terminal ω peptide bond in the *cis* state, rather than the typical *trans* state. The number of trajectories that connected these states was extremely small. Examination of the parallel tempering data revealed that the majority of these transitions had occurred at a much higher temperature, and that the *cis*-ω configurations found at 400 K had reached this temperature by annealing from the higher temperature. In the majority of trajectories at 400 K that contained *cis*-ω configurations, the peptide remained in this state over the duration of the trajectory. This is a clear demonstration of how the automatic algorithm can discover additional slow degrees of freedom that the experimenters may not realize are important. For subsequent investigation, due to the extremely small number of transitions, trajectories containing conformations that included *cis*-ω bonds (a total of 25 trajectories) were removed from the set of trajectories used for analysis, leaving 975 trajectories.

3. Comparison with manual state decomposition

The results of the automatic state decomposition algorithm applied to this reduced data set can be seen in Fig. 3, in comparison with the gold standard manual state decomposition from Ref. 40 and a "poor" manual decomposition that is expected to fail to reproduce kinetics because its states include internal kinetic barriers. (The poor partitioning was defined as follows: (1) $\phi \in [(179,-135\,]$, $\psi \in (98,48]$; (2) $\phi \in (-135,-60]$, $\psi \in (98,48]$; (3) $\phi \in (179,-135]$, $\psi \in (48,98]$; (4) $\phi \in (-135,-60]$, $\psi \in (48,98]$; (5)

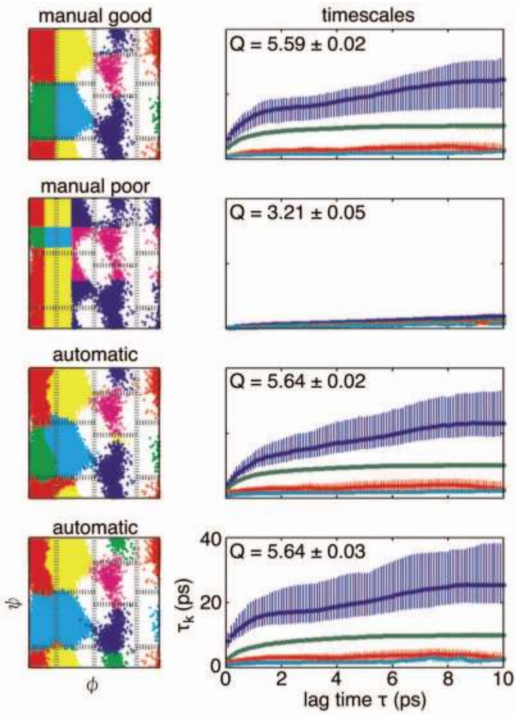

FIG. 3. (Color) Comparison of manual and automatic state decompositions for alanine dipeptide. The left panels depict state partitionings, and the right panels the associated time scales (in picoseconds) as a function of lag time with uncertainties shown, as estimated from the procedure described in Sec. III D. Axes are the same in all plots. Top two panels: Manual "good" or "gold standard" state decomposition from Ref. 40 and manual "poor" state decomposition, where the state boundaries are grossly distorted so as to include internal kinetic barriers within the states. Bottom two panels: Two nearly equivalent partitionings obtained from the automatic state decomposition algorithm.

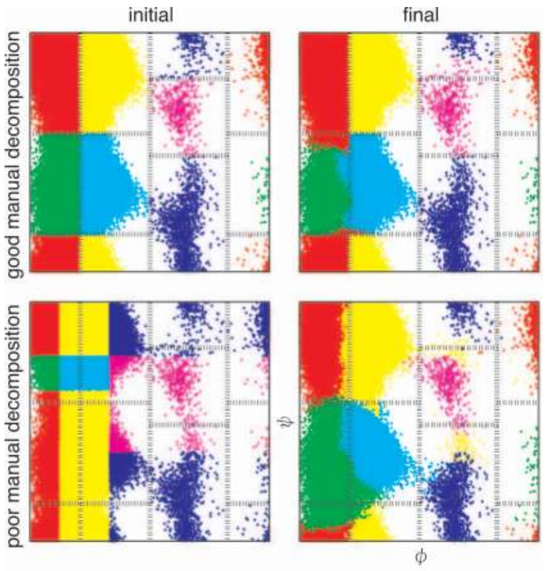

FIG. 4. (Color) Stability and recovery of the optimal state decomposition for alanine dipeptide. Top: Ten cycles of automatic state decomposition applied to a "good" manual partitioning (left) to yield an automatic partitioning (right). Bottom: Ten cycles of automatic state decomposition applied to a "poor" manual partitioning (left) to yield an automatic partitioning (right).

$\phi \in (-60,179]$, $\psi \in (98,-45]$; (6) $\phi \in (-60,179]$, $\psi \in (-45,98]$. Specified intervals denote intervals on the torus, which is continuous from −180 to +180. All torsion angles are specified in degrees.) Independent applications of the automatic method were observed to yield two distinct decompositions with metastabilities within statistical uncertainty, both of which slightly exceeded the metastability of the manual decomposition (Fig. 3, bottom two plots). In the first automatic decomposition, six states in the same general locations as the manual gold standard decomposition are observed, though the boundaries are somewhat perturbed. However, the time scales as a function of lag time are not significantly different from those of the manual gold standard decomposition (Fig. 3, right). In the other automatic decomposition, states 3 and 4 of the manual decomposition (numbering given in Fig. 2) have been merged into a single state, and state 5 of the manual decomposition has been fragmented into two states. Despite this, the time scales as a function of lag time again appear to be statistically indistinguishable from those of the gold standard, suggesting that this model may have equal utility. This suggests that the phenomenological rates may not be very sensitive to the exact choice of state boundaries after the Markov time, as recrossings will have been suppressed by this time. The fact that this lumping does not disrupt the behavior of the model substantially is not altogether surprising, because the barrier separating states 3 and 4 is rather small, and these states act like a single state even on time scales of a few picoseconds or greater. In contrast, the poor decomposition has extremely short time scales which do not appear to level off over the course of 10 ps.

4. Stability of state decomposition

To examine the ability of the algorithm to recover optimal partitionings, the automatic state decomposition algorithm was applied to both the gold standard and poor manual decompositions (Fig. 4) to see whether these partitionings would be maintained over the course of subsequent iterations. Ten iterations were conducted, with each macrostate split to ten microstates in the first iteration, rather than the entire configuration space being split into 100 states. In both cases, the algorithm converged to nearly equivalent partitionings after ten iterations (Fig. 4), as verified by examination of the converged time scales (data not shown). This suggests the method yields partitionings that are relatively stable and optimal. From the poor manual decomposition, however, a number of conformations in manual states 5 and 6 are incorrectly grouped with state 2, though this did not significantly affect the time scales. Further investigation showed that the algorithm never split these conformations from state 2, partly due to their comprising only 1% of the population of the state. Splitting each macrostate into more microstates should alleviate this problem.

B. The F_s helical peptide

To illustrate behavior of the automatic state decomposition method on a larger peptide system with fast kinetics, we applied it to the 21-residue helix-forming F_s peptide, which has been studied extensively both experimentally[76–80] and computationally.[28,81–83] Since helix formation occurs on the

TABLE I. (Color) Macrostates from a 20-state state decomposition of the F_s helical peptide. The backbone is depicted in alpha carbon trace, and arginine sidechains are shown in blue (Arg10), magenta (Arg15), and green (Arg20) for clarity.

state	1	2	3	4	5	6	7
members	358 712	98 222	46 921	22 559	22 367	15 859	11 975
τ_{ac} (ns)	3.1	0.9	1.4	0.6	4.0	1.3	1.6
state	8	9	10	11	12	13	14
members	11 053	11 024	7 976	7 808	7 771	5 978	5 626
τ_{ac} (ns)	2.2	2.0	2.2	1.2	1.6	11.3	2.3
state	15	16	17	18	19	20	
members	4 396	1 856	955	531	525	490	
τ_{ac} (ns)	4.3	5.0	10.3	47.0	29.1	15.2	

nanosecond time scale, Sorin *et al.* were able to reach equilibrium from both helix and coil conformations and observe equilibrium conformational dynamics using ensembles of molecular dynamics trajectories on the distributed computing platform Folding@Home.[28] Two sets of 1000 trajectories at 302 K of varying length of the capped F_s peptide (sequence Ace-A_5[AAARA]$_3$A-Nme), one set initiated from an ideal helix and another from a random coil, were obtained from Sorin *et al.*;[28] details of the simulation protocol are available therein. The first 40 ns of each trajectory, a conservative overestimate of the time to reach equilibrium from either helix or coil, was discarded, and the two sets of trajectories combined to yield a total of 1689 trajectories varying in length from 10 to 95 ns with a sampling interval of 100 ps. In total, this equilibrium data set contained nearly 65 μs of simulation data in 642 604 conformations. The peptide was modeled using the AMBER 99ϕ forcefield[28,84] and solvated in TIP3P water.[74] Though the Berendsen weak-coupling scheme[85] was employed for thermal and pressure control (we note that Berendsen thermal control, here applied independently to the peptide and solvent, modulates the velocities of the peptide atoms during the course of the simulation, which may have a nonphysical effect on dynamics and affect interstate transition rates; however, since we compare our Markov model with the original simulation data set, rather than directly with experiment, this is not of concern), we presume the trajectories still obey microscopic reversibility when only the coordinates of the macromolecular solute are considered for the purposes of computing transition probabilities.

1. Comparison of states

We performed automatic state decomposition on this data set to generate a set of 20 macrostates through ten iterations of splitting and lumping. In the first iteration, the sampled region of conformation space was split into 400 microstates. In subsequent iterations, each macrostate was split into 50 microstates (or, if the expected microstate size was less than 500 configurations, the maximum number of microstates such that the expected microstate size was above 500).

Automatic state decomposition produced a structurally diverse set of states (Table I), ranging in size from over 350 000 to 500 members, with the majority containing from 5000 to 20 000 members. The states include a large state (state 1 of Table I), consisting of slightly over half the total conformations in the data set containing both extended coil and helical conformations; a pure helix state (state 15); a number of helix/coil states which are bent in half to different degrees to form tertiary contacts (states 2-14); and a number of smaller helical states which are bent into circles to form tertiary interactions (states 16-20). A previous analysis[28] of this data clustered conformations into states based on various order parameters: the number of helical residues, number of helical segments (stretches of helical residues), length of the longest helical segment, and radius of gyration. We compared the macrostates generated by the automatic algorithm with these clusters and found that while some states are similar, namely, the binucleated helices of different sizes, most were quite different. The most significant difference was the grouping of helix and coil conformations into a single macrostate in the lumping phase of the automatic algorithm, whereas the order parameter-based clustering kept helix and coil states distinct.[28] When examining individual trajectories, we noticed conformations would rapidly transition between helices and coils between consecutive 100 ps frames of the trajectory, suggesting that their rapid interconversion justifies their lumping into a single macrostate. Additionally, the clustering based on helical order parameters was unable to distinguish certain structures that involved long-lived tertiary contacts, such as the bent and circular helical states. Interestingly, a previous study employing the related AMBER parm03 forcefield[86] identified similar configurations to those noted by the automatic state decomposition, terming these states helix (state 15), helix-turn-helix (states 3, 6–8), adjusted helix-turn-helix (states 4–5, 9–12, 14), and globular helix (states 16–20).

2. Kinetic analysis

We then examined the implied time scales as a function of lag time (Fig. 5). Lumping appeared to preserve the longest time scales found in the microstate transition matrix (data

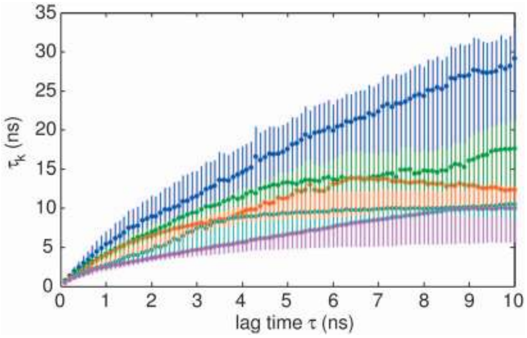

FIG. 5. (Color) Implied time scales of the F_s peptide as a function of lag time for 20-state automatic state decomposition. The five longest time scales are shown. Circles represent the maximum likelihood estimate, and vertical bars depict 68% symmetric confidence intervals about the mean. Note the time scales associated with two processes appear to cross, but are here colored and uncertainties are estimated using the bootstrap procedure by ordering the time scales computed from each bootstrap replicate by rank. This may cause the uncertainties depicted here to be an underestimate of the true uncertainties of each process.

not shown), indicating that our lumping scheme had been successful in identifying a nondestructive lumping into kinetically metastable states at each iteration. Over the course of ten iterations, the metastability (as optimized with a lag time of 100 ps) increased from 12.5±0.3 to 14.5±0.1, suggesting that the iterative refinement was improving the quality of the state decomposition. On the first iteration, the longest time scales increase nearly linearly with lag time, while on the last iteration, some of the longest time scales become stable by a lag time of 4–5 ns, suggesting Markovian behavior for some of the processes.

Using the interpretation of eigenvector components in terms of aggregate modes described in Sec. II A, the longest time scale was found to correspond to movement between the extended helix/coil state (state 1) and one of the twisted helix-turn-helix states (state 18) with only 500 members. We found, however, that state 18 appeared a small number of times in 30 trajectories, and over 450 times in a single trajectory. Further examination revealed that conformations belonging to this state were almost exclusively temporally adjacent to conformations belonging to state 5, and structural comparison of conformations of these two states showed they were strikingly similar. This suggests that slight conformational differences between conformations in states 18 and 5 allowed the K-medoid clustering algorithm to partition between these states in a splitting step, and since state 18 was mainly isolated in a single trajectory, its self-transition probability was maximized by *not* lumping it with state 5, even though the two behaved in a similar kinetic fashion. Indeed, when we manually lump states 18 and 5, the longest time scale, corresponding to transitions involving state 18, disappears, but the remaining time scales are all preserved (data not shown).

A potential cause of the increase with lag time observed in some of the other long time scales may be due to the finite length of trajectories. If the state is long lived and occurs near the trajectory beginning or end, then it can be seen that the estimated self-transition probability T_{ii} artificially increases as a function of lag time. This effect is most pronounced when a state occurs in very few trajectories, and appears to be mitigated when the state occurs in many trajectories at random times within the trajectory.

In order to determine which states are poorly characterized, we estimated the number of statistically independent visits to each macrostate using the autocorrelation time given in Sec. III D. As the correlation functions became statistically unreliable at times larger than 10 ns, a least squares linear fit to the log of the computed correlation function over the first 10 ns was used to estimate the tail at times greater than 10 ns, and this combined correlation function was integrated to obtain the autocorrelation time. Computed state autocorrelation times are given in Table I. For many states, the correlation time was 1–2 ns, giving thousands of independent samples; however, for five states, including the four involved in the four longest time scales, the correlation times were between 10 and 50 ns, suggesting that the data set contained less than 50 independent samples of these states. Currently, in the automatic state decomposition algorithm, we try to reduce the statistical uncertainty in the transition matrix by limiting the expected population of each state to be greater than some minimum number of configurations. Since the conformations appearing within some states may be highly correlated, the number of conformations within a state is not the best measure of how statistically well determined its transition elements are; instead, it may be advantageous to place a lower limit on the effective number of independent visits to each state, which is far less than the number of configurations it contains. Alternatively, it may be necessary to ensure better characterization of these states by conducting additional simulations from them, provided the equilibrium transition probabilities can still be computed.

We constructed a Markov model from the transition matrix estimated at a 5 ns lag time, where some (though not all) of the time scales appear to have stabilized. The Chapman-Kolmogorov test (Sec. II D 2) can assess how well the model reproduces the observed kinetics. The time evolution of probability density out of three states (state 2, a populous state; state 13, a moderately populated state; and state 19, a sparsely populated state) over the course of 50 ns is shown in Fig. 6. The Markov model appears to do a very reasonable job of predicting the time evolution of the system to within statistical uncertainty over many times longer than the lag time used to construct it. In fact, the time evolution was well modeled for evolution out of all states, except for state 13, for which dynamics seemed to be particularly poorly reproduced. This state has a long correlation time, and many trajectories seem to contain only a single configuration that is part of this state, suggesting its boundaries are simply poorly resolved. Regardless, the time evolution is generally well modeled for this system.

C. The trpzip2 β-peptide

As an illustration of the application of the state decomposition algorithm to a system with complex kinetics implying the existence of multiple metastable states,[87] we considered the engineered 12-residue β-peptide trpzip2.[88] A set of 323 10 ns constant-energy, constant-volume simulations of

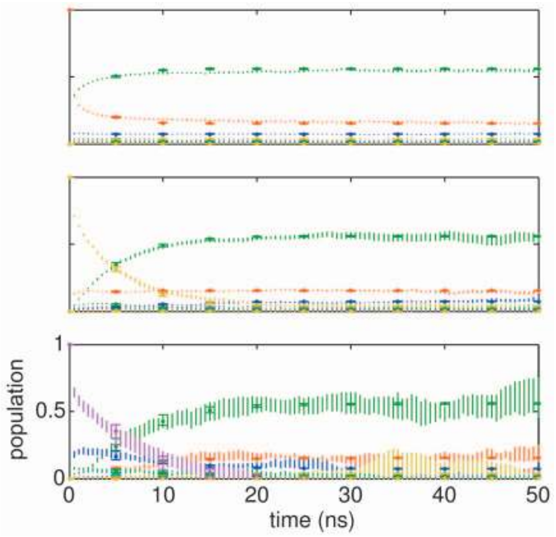

FIG. 6. (Color) Reproduction of observed state population evolution by a Markov model for the F_s peptide. The time evolution of the Markov model constructed from the 5 ns lag time transition matrix is shown by the filled circles with flat error bars, which denote the 68% confidence interval estimated from a sample of 40 bootstrap realizations, with each realization the result of a new transition matrix estimated from a bootstrap sample of trajectories. Vertical bars without flat ends denote the 68% confidence interval centered on the sample mean for the probability of finding the system in the 20 macrostates a given time after initial preparation in a specific state. The system was originally prepared in state 2 (top, red), 13 (middle, yellow), or 19 (bottom, purple). The most populous states are colored green (state 1), red (state 2), and blue (state 3).

the unblocked peptide (note that the peptide studied experimentally in Refs. 87 and 88 was synthesized with an amidated C terminus, whereas the termini of the simulated peptide in the data set considered here were left zwitterionic) simulated using the AMBER 96 forcefield[72] in TIP3P water[73] was obtained from Pitera *et al.*,[89] details of the simulation protocol are provided therein. The trajectories were initiated from an equilibrium sampling of configurations at 425 K, a temperature high enough to observe repeated unfolding and refolding events at equilibrium. Configurations were sampled every 10 ps, giving a total of 3.23 μs of data in 323 000 configurations.

1. Comparison of states

The automatic state decomposition method was applied to obtain a set of 40 macrostates in ten iterations of splitting and lumping. The algorithm was performed as described in Sec. III C, except for the first iteration, where the conformations were split into 400 microstates.

Figure 7 depicts some of the final set of 40 macrostates compared with a set of states identified by consideration of backbone hydrogen bonding patterns in the previous study by Pitera *et al.*[89] (The complete set of macrostates is shown in a figure included in Ref. 90.) As the trajectories considered here were resampled to 10 ps intervals (rather than 1 ps as in Ref. 89), we found less than five examples of the +2 and −2 hydrogen bonding states identified in Ref. 89, and therefore exclude them from comparison. The automatic state decomposition method recovers states corresponding to the native, +1C, and +1N hydrogen bonding patterns, and often further resolves them based on the orientation of the tryptophan side chains (Figs. 7, A, C, and D). However, the −1N hydrogen bonding pattern is not further resolved, and instead is grouped into a state of mostly disordered hairpins; further examination is necessary to determine whether the algorithm simply failed to resolve this state or if the state is simply not long lived. In addition to recovering most of the manually identified misregistered states, the algorithm was

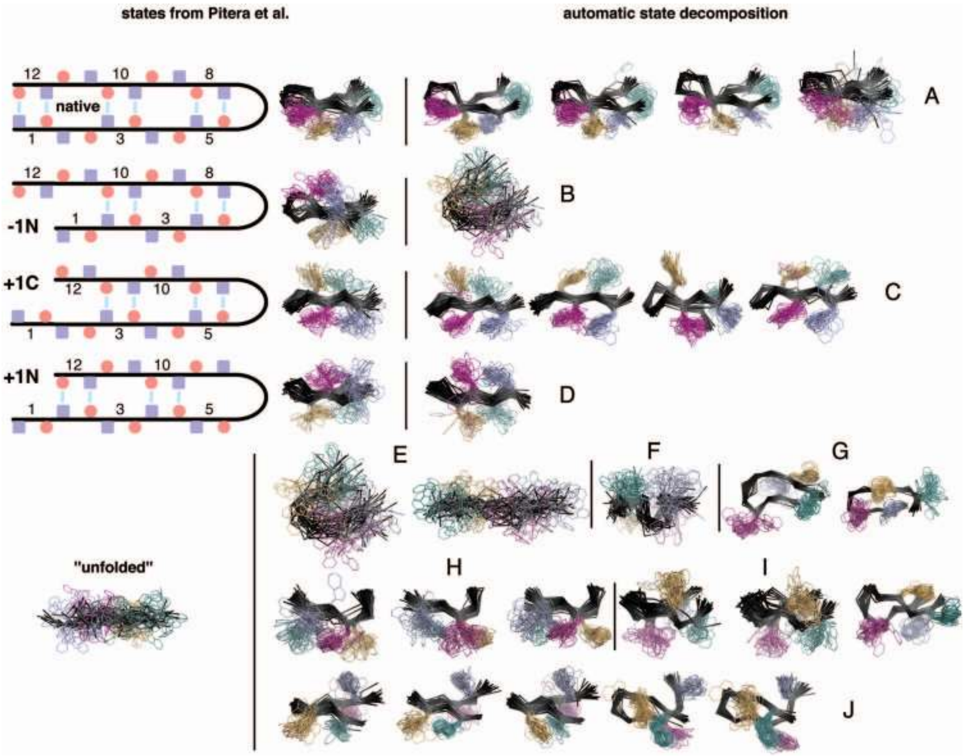

FIG. 7. (Color) Comparison of some trpzip2 macrostates found by automatic state decomposition with misregistered hydrogen bonding states identified in a previous study. Left: The five hydrogen bonding patterns enumerated in Pitera *et al.* (Ref. 89) that occurred in sufficient numbers in the subsampled trpzip2 data set used here, with representative conformational ensembles. Blue squares denote backbone amide hydrogen bond donors, and red circles denote backbone carbonyl hydrogen bond acceptors. Right: A selection of macrostates discovered by automatic state decomposition that contain the largest numbers of hydrogen bonding pattern states. The backbone is depicted in alpha carbon trace, and tryptophan side chains are shown in light blue (Trp2), orange (Trp4), magenta (Trp9), and teal (Trp11). A complete set of macrostates obtained from the 40-state decomposition of the trpzip2 data set is available as supplementary information.

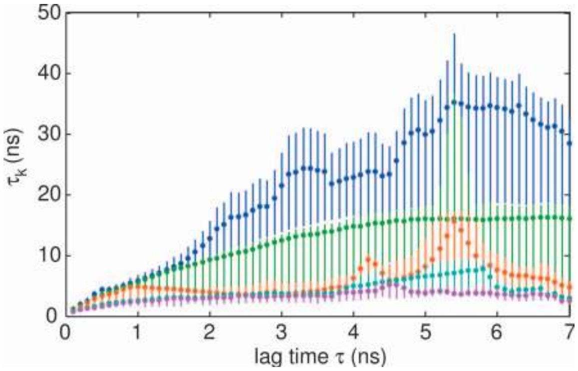

FIG. 8. (Color) Implied time scales of trpzip2 as a function of lag time for 40-state automatic state decomposition. The five longest time scales are shown.

also able to greatly resolve the state labeled as "unfolded" in Pitera *et al.*[89] (in that it did not conform to any of the enumerated hydrogen bonding patterns) into substates which exhibit considerable structure (E-J). Some of these kinetically resolved states have distinct hydrogen bonding patterns, such as where both strands are rotated (H), causing the tryptophan side chains to appear on the opposite face, or where the misregistration is greater than two residues (G and J). This demonstrates the utility of the method in identifying additional kinetically relevant states that were not initially part of the experimental hypothesis space.

2. Kinetic analysis

Figure 8 depicts the implied time scales of the kinetic model as a function of lag time. The longest time scale ranges between 25 and 35 ns and appears to stabilize over the range of lag times considered, though the uncertainty is quite large. Eigenvector analysis (described in Sec. II A) shows that this time scale corresponds to transitions between the unfolded and disordered hairpin states (E) and the hairpin with both strands rotated (H). The states labeled H together totaled 935 conformations, but appeared in only 13 trajectories, with over 95% of the conformations appearing in a single trajectory. Correlation time analysis (Sec. III D) suggests there are less than ten independent samples for each of the three states, so proper resolution of this time scale would require more data. The second longest time scale grows to about 15 ns, levels off by around 4 ns, and corresponds to transitions between the unfolded and disordered hairpin states (E) and the native backbone states (A). The states involved in this transition are much better characterized, with a total of over 25 000 conformations appearing in over half the trajectories. The next three longest time scales were all between 3 and 4 ns and correspond to movement between the unfolded state (E) and various sets of misregistered states, namely, the newly identified misregistered states I and J, and the +1C state (C). Unfortunately, these time scales are on the order of the time to reach global equilibrium, so it is difficult to characterize these transitions well.

V. DISCUSSION

Markov models are expected to be effective and efficient ways to statistically summarize information about the pathways (mechanism) and time scales for heterogeneous biomolecular processes such as protein folding. The great challenge in their use lies in defining an appropriate state space. Here, we have presented a new algorithm for automatically generating a set of configurational states that is appropriate for describing peptide conformational dynamics in terms of a Markov model, though we expect it to be applicable to macromolecular dynamics in general. The algorithm uses molecular dynamics simulations as input, and generates state definitions using information about the temporal order of conformations seen in the trajectories. The importance of having an automatic algorithm, i.e., one that requires little or no human intervention, is that without it, human bias may inadvertently produce incorrect interpretations of the mechanism of conformational change by imposing a particular view of the simulation data. Additionally, molecular simulation data sets are becoming so large and complex that effectively summarizing the data or extracting insight becomes increasingly impractical unless the experimenter analyzes the data with a specific hypothesis in mind. Construction of a Markov model, however, allows for a "hypothesis-free" investigation of conformational dynamics, provided that the state space is sufficiently well sampled.

Our algorithm is based on the availability of large numbers of molecular dynamics simulations of appropriate simulation length such as might be generated by a supercomputer or a large (possibly distributed) cluster. Current technology allows for the production of thousands of simulations that can be tens of nanoseconds in length, hundreds of trajectories of up to hundreds of nanoseconds in length, or dozens that are on the order of a microsecond in length. Since our goal has been to develop Markov models that accurately characterize the time evolution of ensembles of macromolecules over experimental time scales (that can range from microseconds to milliseconds) from short simulations of single molecules, our approach places strong emphasis on the longest time scales observed in molecular simulations. For example, recognizing that ill-formed states often result in artificially shortened time scales, we sought to find states that maximize the time scales implied by their corresponding transition matrix for a particular choice of lag time and number of states. This resulted in the maximization of the metastability as a computationally convenient surrogate for minimizing the internal equilibration time τ_{int}.

For the three data sets to which we have applied the method, there have been a number of important successes. For alanine dipeptide, the algorithm discovered a distinct manifold of states that consisted of conformations containing a *cis*-ω peptide bond. This manifold was discovered because it was kinetically distinct, rather than structurally distinct. Also, for alanine dipeptide, the method produces states that are robust and structurally very similar to the best ones produced manually, as well as kinetically indistinguishable to within statistical uncertainty according to our validation metrics. The application of the method to the F_s peptide data set produced a set of states somewhat different from those identified previously from the clustering of helical order parameters.[28] The states produced by the algorithm properly identified many very long lived (metastable) conformations whose lifetimes and kinetics might be experimentally relevant. The Markov model produced from this state decomposition and a 5 ns transition matrix was shown to reproduce the observed state populations over 50 ns to within statistical uncertainty. Finally, for the application of the method to the trpzip2 peptide the states constructed were consistent with ones previously identified.[89] This was very encouraging since the previously constructed states used an intramolecular hydrogen bonding criterion and the automatic algorithm utilized different observables and metrics, heavy atom RMSD and kinetics, to resolve states. Moreover, the automatic algorithm more finely resolved what was considered to be the unfolded ensemble into metastable states that were not identified by the decomposition based on hydrogen bonding patterns.

Therefore, the algorithm is achieving many of its design objectives. It provides a method for identifying and characterizing the *slower* degrees of freedom of a molecular system. It correctly identifies metastable states, dividing structurally similar conformations into multiple sets that have short times for intraconversion but long times for interconversion, and combines conformations that rapidly interconvert even though they may be structurally diverse. This is a prerequisite to capturing a concise description of the pathways for conformational changes. Once meaningful states are identified, the transition matrix itself encapsulates the branching ratios for various pathways and the time scales for overall relaxation to equilibrium from any arbitrary starting ensemble.

Work is ongoing to establish standards for the amount and nature of simulation data (number and length of simulations) needed to develop useful and sufficiently precise Markov models as well as investigations of the effect of quality metrics other than the metastability on the nature of the resulting states and time scales. Metrics for assessing the quality of the resulting model also need to be examined to complement, or as alternatives to, seeking stability of the implied time scales with respect to lag time. Finally, alternative approaches to performing this state decomposition are a further matter of current study, such as the method of Noé *et al.* appearing in this issue, motivated by much the same ideas of metastability but employing different methods for the construction of a microstate space.[91]

A general observation about the models produced using states defined by our method is that Markovian behavior is not obtained until lag times that are only an order of magnitude shorter than the longest time scales. Recall that the *utility* of a state space depends to a large extent on how early Markovian behavior is observed compared to the processes of interest. There are multiple possibilities for why this might be the case. For some molecular systems, there may be no identifiable metastable states in the usual sense. The existence of experimentally observed metastable states in protein systems (e.g., native, intermediate, and unfolded) combined with the observation of metastable states in models of small solvated peptides[40] argues that this is unlikely. It could be that statistical uncertainty is undermining both the metastability quality metric and the tests for Markovian behavior. Alternatively, the way we establish boundaries between states may not be flexible enough to adequately divide true metastable regions. It may also be that we simply need to allow more states to be produced, resulting in subdivision of states that have internal barriers, to reduce the Markov times. Both of these latter possibilities could in principle be easily addressed by allowing the creation of more states. However, the creation of more states, especially ones with low populations, leads inevitably to situations where transition probabilities become statistically unreliable given a fixed quantity of equilibrium data.

Long time scales are ultimately the result of infrequent events, and even for large but finite equilibrium data sets these will be small in number, with resulting small off-diagonal transition probabilities that are statistically unreliable. This has placed us in the particularly difficult but unavoidable situation of attempting to optimize a statistically uncertain objective function. One solution to this problem, of course, is to consider this algorithm as only the first step of an iterative process where important states and transitions are identified, and then further simulations are performed to improve the characterization of important regions of conformation space. This will allow refinement of the state space and improved precision for important selected transition probabilities. Information from the subsequent simulations could be combined with that from the first set using the selection cell approach described previously.[39] Selection of states, or regions of configuration space, from which further simulations should be initiated could be chosen based on uncertainty considerations.[31]

ACKNOWLEDGMENTS

The authors would especially like to thank Jed W. Pitera (IBM) for insightful discussion and constructive comments on this manuscript and for providing simulation data for trpzip2, Eric Sorin (Stanford) for providing simulation data for the F_s peptide, Hans C. Andersen (Stanford) and Frank Noé (IWR Heidelberg) for enlightening conversations on the nature of Markov chain models, Vishal Vaidyanathan (Stanford) for assistance with clustering algorithms, and Libusha Kelly and David L. Mobley (UCSF) and Guha Jayachandran (Stanford) for critical comments on this manuscript. One of the authors (J.D.C.) was supported by a Howard Hughes Medical Institute and an IBM predoctoral fellowship. Another author (W.C.S.) acknowledges support from NSF MRSEC Center on Polymer Interfaces and Macromolecular Assemblies DMR-0213618. Another author (K.A.D.) acknowledges the support of NIH Grant No. GM34993. Two of the authors (N.S. and V.S.P.) acknowledge support from NSF Grant No. 0317072. Two of the authors (J.D.C. and N.S.) contributed equally to this work.

[1] C. M. Dobson, Nature (London) **426**, 884 (2003).
[2] E. Z. Eisenmesser, D. A. Bosco, M. Akke, and D. Kern, Science **295**, 1520 (2002).
[3] B. Youngblood and N. O. Reich, J. Biol. Chem. **281**, 26821 (2006).

[4] D. D. Boehr, D. McElheny, H. J. Dyson, and P. E. Wright, Science **313**, 1638 (2006).

[5] H. Frauenfelder, B. H. McMahon, R. H. Austin, K. Chu, and J. T. Groves, Proc. Natl. Acad. Sci. U.S.A. **98**, 2370 (2001).

[6] J. Changeux and S. J. Edelstein, Science **308**, 1424 (2005).

[7] N. Maki, K. Moitra, P. Ghosh, and S. Dey, J. Biol. Chem. **281**, 10769 (2006).

[8] D. Moroni, T. S. van Erp, and P. G. Bolhuis, Physica A **340**, 395 (2004).

[9] A. K. Faradjian and R. Elber, J. Chem. Phys. **120**, 10880 (2004).

[10] Y. M. Rhee and V. S. Pande, J. Phys. Chem. B **109**, 6780 (2005).

[11] A. Berezhkovskii and A. Szabo, J. Chem. Phys. **122**, 014503 (2005).

[12] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry*, 2nd ed. (Elsevier, New York, 1997).

[13] R. Czerminski and R. Elber, J. Chem. Phys. **92**, 5580 (1990).

[14] R. E. Kunz and R. S. Berry, J. Chem. Phys. **103**, 1904 (1995).

[15] K. D. Ball and R. S. Berry, J. Chem. Phys. **109**, 8557 (1998).

[16] Y. Levy, J. Jortner, and O. M. Becker, J. Chem. Phys. **115**, 10533 (2001).

[17] P. N. Mortenson and D. J. Wales, J. Chem. Phys. **114**, 6443 (2001).

[18] P. N. Mortenson, D. A. Evans, and D. J. Wales, J. Chem. Phys. **117**, 1363 (2002).

[19] D. A. Evans and D. J. Wales, J. Chem. Phys. **112**, 1080 (2004).

[20] D. Shalloway, J. Chem. Phys. **105**, 9986 (1996).

[21] A. Ulitsky and D. Shalloway, J. Chem. Phys. **109**, 1670 (1998).

[22] M. Shen and K. F. Freed, J. Chem. Phys. **118**, 5143 (2003).

[23] H. Grubmüller and P. Tavan, J. Chem. Phys. **101**, 5047 (1994).

[24] B. L. de Groot, X. Daura, A. E. Mark, and H. Grubmüller, J. Mol. Biol. **309**, 299 (2001).

[25] W. C. Swope, J. W. Pitera, F. Suits *et al.*, J. Phys. Chem. B **108**, 6582 (2004).

[26] N. Singhal, C. D. Snow, and V. S. Pande, J. Chem. Phys. **121**, 415 (2004).

[27] M. Andrec, A. K. Felts, E. Gallicchio, and R. M. Levy, Proc. Natl. Acad. Sci. U.S.A. **102**, 6801 (2005).

[28] E. J. Sorin and V. S. Pande, Biophys. J. **88**, 2472 (2005).

[29] S. Sriraman, I. G. Kevrekidis, and G. Hummer, J. Phys. Chem. B **109**, 6479 (2005).

[30] V. Schultheis, T. Hirschberger, H. Carstens, and P. Tavan, J. Chem. Theory Comput. **1**, 515 (2005).

[31] N. Singhal and V. S. Pande, J. Chem. Phys. **123**, 204909 (2005).

[32] S. P. Elmer, S. Park, and V. S. Pande, J. Chem. Phys. **123**, 114903 (2005).

[33] S. Park and V. S. Pande, J. Chem. Phys. **124**, 054118 (2006).

[34] D. Chandler, J. Chem. Phys. **68**, 2959 (1978).

[35] A. Ansari, J. Berendzen, S. F. Bowne, H. Frauenfelder, I. E. T. Iben, T. B. Sauke, E. Shyamsunder, and R. D. Young, Proc. Natl. Acad. Sci. U.S.A. **82**, 5000 (1985).

[36] Y. S. Bai and M. D. Fayer, Phys. Rev. B **39**, 11066 (1989).

[37] O. M. Becker and M. Karplus, J. Chem. Phys. **106**, 1495 (1997).

[38] Y. Levy, J. Jortner, and R. S. Berry, Phys. Chem. Chem. Phys. **4**, 5052 (2002).

[39] W. C. Swope, J. W. Pitera, and F. Suits, J. Phys. Chem. B **108**, 6571 (2004).

[40] J. D. Chodera, W. C. Swope, J. W. Pitera, and K. A. Dill, Multiscale Model. Simul. **5**, 1214 (2006).

[41] S. B. Ozkan, K. A. Dill, and I. Bahar, Protein Sci. **11**, 1958 (2002).

[42] P. Lenz, B. Zagrovic, J. Shapiro, and V. S. Pande, J. Chem. Phys. **120**, 6769 (2004).

[43] M. E. Karpen, D. J. Tobias, and C. L. Brooks III, Biochemistry **32**, 412 (1993).

[44] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, J. Comput. Phys. **151**, 146 (1999).

[45] C. Schütte and W. Huisinga, in *Handbook of Numerical Analysis: Special Volume on Computational Chemistry*, edited by P. G. Ciaret and J.-L. Lions (Elsevier, New York, 2002), Vol. X.

[46] *Stochastic Processes in Chemical Physics: The Master Equation*, edited by I. Oppenheim, K. E. Shuler, and G. H. Weiss (MIT, Cambridge, MA, 1977).

[47] C. Schütte, Ph.D. thesis, Konrad Zuse Zentrum Berlin, 1999.

[48] W. Huisinga, Ph.D. thesis, Free University of Berlin, 2001.

[49] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Clarendon, Oxford, 1991).

[50] M. Weber, Ph.D. thesis, Free University of Berlin, 2006.

[51] E. Meerback, C. Schütte, and A. Fischer, Linear Algebr. Appl. **398**, 141 (2005).

[52] J. D. Chodera, W. C. Swope, J. W. Pitera, and K. A. Dill (unpublished).

[53] B. G. Fitch, R. S. Germain, M. Mendell *et al.*, J. Parallel Distrib. Comput. **63**, 59 (2003).

[54] R. S. Germain, B. Fitch, A. Rayshubskiy, M. Eleftheriou, M. C. Pitman, F. Suits, M. Giampapa, and T. C. Ward, CODES+ISSS *'05*, Proceedings of the 3rd IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis, Jersey City, NJ, 2005 (ACM Press, New York, NY) pp. 207–212.

[55] M. Shirts and V. S. Pande, Science **290**, 1903 (2000).

[56] V. S. Pande, I. Baker, J. Chapman *et al.*, Biopolymers **68**, 91 (2003).

[57] W. Huisinga and B. Schmidt, *New Algorithms for Macromolecular Simulation*, Lecture Notes in Computational Science and Engineering Vol. 49 (Springer, New York 2006), Part III.

[58] D. G. Truhlar, B. C. Garrett, and S. J. Klippenstein, J. Phys. Chem. **100**, 2771 (1996).

[59] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer, New York, 2001).

[60] J. MacQueen, *Proceedings of the Fifth Berkeley Symposium on Mathematical statistics and probability* (University of California Press, Berkeley, 1967) pp. 281–297.

[61] D. L. Theobald, Acta Crystallogr., Sect. A: Found. Crystallogr. **A61**, 478 (2005).

[62] B. Steipe, Acta Crystallogr., Sect. A: Found. Crystallogr. **A58**, 506 (2002).

[63] P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte, Numer. Linear Algebra Appl. **315**, 39 (2000).

[64] B. Efron, Ann. Stat. **7**, 1 (1979).

[65] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson, J. Chem. Phys. **76**, 637 (1982).

[66] W. Janke, in *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms*, Vol. 10, edited by J. Grotendorst, D. Marx, and A. Murmatsu (John von Neumann Institute for Computing, Jülich, Germany, 2002), pp. 423–445.

[67] J. Apostolakis, P. Ferrara, and A. Caflisch, J. Chem. Phys. **110**, 2099 (1999).

[68] P. G. Bolhuis, C. Dellago, and D. Chandler, Proc. Natl. Acad. Sci. U.S.A. **97**, 5877 (2000).

[69] G. Hummer and I. G. Kevrekidis, J. Chem. Phys. **118**, 10762 (2003).

[70] D. S. Chekmarev, T. Ishida, and R. M. Levy, J. Phys. Chem. B **108**, 19487 (2004).

[71] A. Ma and A. R. Dinner, J. Phys. Chem. B **109**, 6769 (2005).

[72] P. A. Kollman, R. Dixon, W. Cornell, T. Vox, C. Chipot, and A. Pohorille, in *Computer Simulation of Biomolecular Systems*, edited by A. Wilkinson, P. Weiner, and W. F. van Gunsteren (Kluwer/Escom, Leiden, The Netherlands, 1997), Vol. 3, pp. 83–96.

[73] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).

[74] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, J. Comput. Chem. **13**, 1011 (1992).

[75] J. D. Chodera, W. C. Swope, J. W. Pitera, C. Seok, and K. A. Dill, J. Chem. Theory Comput. **3**, 26 (2007).

[76] D. J. Lockhart and P. S. Kim, Science **257**, 947 (1992).

[77] D. J. Lockhart and P. S. Kim, Science **260**, 198 (1993).

[78] S. Williams, T. P. Causgrove, R. Gilmanshin, K. S. Fang, R. H. Callender, W. H. Woodruff, and R. B. Dyer, Biochemistry **35**, 691 (1996).

[79] P. A. Thompson, W. A. Eaton, and J. Hofrichter, Biochemistry **36**, 9200 (1997).

[80] I. K. Lednev, A. S. Karnoup, M. C. Sparrow, and S. A. Asher, J. Am. Chem. Soc. **123**, 2388 (2001).

[81] A. E. García and K. Y. Sanbonmatsu, Proc. Natl. Acad. Sci. U.S.A. **99**, 2782 (2002).

[82] W. Zhang, H. Lei, S. Chowdbury, and Y. Duan, J. Phys. Chem. B **108**, 7479 (2004).

[83] E. J. Sorin and V. S. Pande, J. Comput. Chem. **26**, 682 (2005).

[84] J. Wang, P. Cieplak, and P. A. Kollman, J. Comput. Chem. **21**, 1049 (2000).

[85] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, J. Chem. Phys. **81**, 3684 (1984).

[86] Y. Duan, C. Wu, S. Chowdhury *et al.*, J. Comput. Chem. **24**, 1999 (2003).

[87] W. Y. Yang and M. Gruebele, J. Am. Chem. Soc. **126**, 7758 (2004).

[88] A. G. Cochran, N. J. Skelton, and M. A. Starovasnik, Proc. Natl. Acad. Sci. U.S.A. **98**, 5578 (2001).

[89] J. W. Pitera, I. Haque, and W. C. Swope, J. Chem. Phys. **124**, 141102

(2006).

[90] A FORTRAN 90/95 implementation of the automatic state decomposition algorithm presented here is available for download as part of the supplementary information for this article. The latest version of the code, along with the alanine dipeptide data set, can be obtained from http://www.dillgroup.ucsf.edu/~jchodera/code/automatic-state-decomposition/. The trpzip2 data set is available directly from W. C. Swope upon request (electronic mail: swope@almaden.ibm.com). A gallery of all macrostates produced by the 40-state decomposition of the trpzip2 peptide is also available as part of the supplementary information for this article.

[91] F. Noé, I. Horenko, C. Schütte, and J. C. Smith, J. Chem. Phys. **126**, 155102 (2007).