Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory†

William C. Swope* and Jed W. Pitera

IBM Almaden Research Center, 650 Harry Road, San Jose, California 95120

Frank Suits

IBM Watson Research Center, Route 134, Yorktown Heights, New York 10598

Received: November 10, 2003; In Final Form: February 21, 2004

A rigorous formalism for the extraction of state-to-state transition functions from a Boltzmann-weighted ensemble of microcanonical molecular dynamics simulations has been developed as a way to study the kinetics of protein folding in the context of a Markov chain. Analysis of these transition functions for signatures of Markovian behavior is described. The method has been applied to an example problem that is based on an underlying Markov process. The example problem shows that when an instance of the process is analyzed under the assumption that the underlying states have been aggregated into macrostates, a procedure known as lumping, the resulting chain appears to have been produced by a non-Markovian process when viewed at high temporal resolution. However, when viewed on longer time scales, and for appropriately lumped macrostates, Markovian behavior can be recovered. The potential for extracting the long time scale behavior of the folding process from a large number of short, independent molecular dynamics simulations is also explored.

1. Introduction

An understanding of the mechanisms by which proteins fold would have wide utility in many areas, ranging from the development of effective treatments for protein folding related diseases to exploitation of the underlying principles of folding to facilitate industrial nanotechnology. The study of protein folding has three aspects: thermodynamics, kinetics, and structure prediction. In this work we introduce an approach to characterizing some aspects of protein folding kinetics and apply it to a simple example problem. In a companion paper,[1] we apply the approach to the folding of a small peptide, the C-terminal β-hairpin motif from protein G.

Protein folding has been extensively studied experimentally[2−6] and by computer simulation.[7−12] Computer simulations can provide information about the process that is highly complementary to that obtained from experiment.[8,13−17] Furthermore, the computer power available for biomolecular simulations in general, and protein folding in particular, is increasing through the production of improved software to exploit parallelism,[18] specialized hardware,[19] larger and faster computer systems and grid and distributed computing approaches.[20−23] Indeed, the IBM BlueGene project,[24−27] to build a massively parallel computer to investigate biomolecular processes such as protein folding, is expected to systematically study a variety of peptide and small protein systems and will produce very large volumes of simulation data. One significant advantage of this greater computer power is that the field is moving from studies that report on single events observed during single trajectories of limited duration,[7] to studies where extensive thermodynamic sampling has been performed[11−13,28−30] and ensembles of trajectories are produced and analyzed.[8,9] Obtaining large numbers of independent trajectories is not only a very effective way to use parallel computing technologies but is required for statistically meaningful and reproducible results.[31] Because of this move to more comprehensive simulations, new and automatable analysis procedures that can be applied consistently to data from simulations of a variety of protein systems need to be developed and validated.

Protein folding is generally studied in the liquid phase, where the protein or peptide is in contact with a solvent. Besides providing part of the driving force for the folding process, through hydrophobic and hydrophilic hydration, the solvent also provides friction and a heat bath for the process. In fact, because of the random forces exerted by the solvent, one would expect that if several identical peptides could be prepared in the same conformation and solvated, they would very likely adopt different folding trajectories, perhaps following completely different paths and taking different amounts of time to reach native conformations. It is because of this stochastic nature of folding that one should be careful not to draw strong conclusions about the process if they are deduced from single MD trajectories. But given that hundreds of protein simulation trajectories can be produced, what is the best way to use them to understand the process of folding? One possible approach, explored in this work, is to analyze the trajectories to produce a *probability* for the evolution of the protein from one conformational state to another. The formalism associated with Markov processes and models is, therefore, a natural approach for this analysis.

Markov models of stochastic processes deal with the temporal evolution of the state of a system. They are appropriate when the memory of the system is short. That is, when the evolution of the system into the near future depends only its properties at the current time, and not on any of its prior history. Markov models can be of several types depending on whether one discretizes the time domain, the state space, or both. With a discrete time Markov chain, both the time and space domain

† Part of the special issue "Hans C. Andersen Festschrift".

* To whom correspondence should be addressed.

10.1021/jp037421y CCC: $27.50 © 2004 American Chemical Society
Published on Web 04/22/2004

are discretized. These types of models of stochastic processes involve a finite time transition matrix that is capable of describing the long time behavior of the system. This matrix gives the probability that during some discrete interval of time the system makes transitions between *states*. The matrix is best regarded as a propagator for a probability distribution. Markov models of physical processes are quite common and useful in chemical physics.[45] If protein folding is a physical process for which a Markov description is appropriate, it could be possible to characterize the long time behavior of a solvated protein system through a number of relatively short time molecular dynamics simulations.

Protein folding kinetics, in this view, is about the evolution of a probability density for an ensemble of proteins as it relaxes from some nonequilibrium to an equilibrium distribution. This view has close connection with the experimental measurement of folding rates, because many of the experiments observe the temporal evolution of some spectroscopic signal as an *ensemble* of proteins evolves to equilibrium after a thermodynamic perturbation such as heating or cooling, addition of denaturant, changes in pressure, etc.

Theoretical approaches to protein folding that view the process in the context of Markov models are not new. Earlier work[32,33] on lattice models of proteins have used an ansatz for the Markov transition rates between states as a means of relating the evolution of an ensemble of lattice conformations to experimental rates. Recent work by Ozkan[33] and by Zhang and Chen[34] describe ways to deduce characteristics of folding pathways, transition states**,** and long-lived intermediates from an analysis of the eigenvalues and eigenvectors of the transition matrix itself, and they have also applied them to cases where an ansatz was used for the transition rates between pairs of states. If stable and metastable states, as well as transition states, of a system are known a priori, or can be guessed, it is also common to extract or estimate Markov transition rates between these states through the use of simulation.[35−38]

To model a physical process with a Markov chain approach, an appropriate state space needs to be established. However, it is far from obvious in general how one should tabulate and characterize the stable and transition states of a protein system. In fact, doing so predetermines the outcome of the analysis to a great extent. Moreover, an incorrect choice for the state space can make a Markovian analysis inappropriate. An ideal method would be able to construct an appropriate state space without assuming prior knowledge of the existence or nature of those states.

Regarding the construction of appropriate states, it is important to recognize that although the underlying classical dynamics of an energy conserving and time reversible MD simulation is inherently a Markov process, the formulation of states that aggregate *finite* regions of phase space can result in behavior that defies a Markovian description. In fact, the nature of processes that can be described as Markov chains in some state space, but are viewed in a different state space that consists of aggregated, or *lumped*, versions of the original states, is an area of much current study.[39−41] In particular, what are the characteristics of partitionings of states that preserve the Markovian behavior in the reduced, or aggregated, state space?

Because the goal of a Markov description is to extract accurate temporal behavior, it is important that the states are defined in a kinetically meaningful way. This imposes several requirements on the states. First, trajectories that pass through the phase space spanned by any one of these states should behave in a *dynamically* similar way. Second, because it is natural to define states with respect to order parameters related to the folding process, and to establish states that are compact in this order parameter space, the order parameters themselves must satisfy a "kinetic ruler" characteristic:[32,42] temporal progress during, e.g., typical folding trajectories should correspond to monotonic (preferably linear) changes in the order parameter. Order parameters in common use that may be appropriate for characterizing thermodynamics, such as fraction of native contacts formed, are probably not appropriate for kinetic studies, because topologically very different conformations may have the same value for this order parameter, and these conformations are likely to exhibit very different kinetic behavior. In fact, it is not obvious how to select appropriate order parameters for kinetics, but it is probably true that a Markov description of the process will not be possible without them.

Furthermore, it is possible that protein folding is not truly Markovian on the time scales that are accessible within MD simulations. Even for a good choice of state space, for a Markov description of the process to be accurate, there is a minimum time interval over which transitions within the system can be described by a history-independent transition matrix. This time interval, below which transitions between states will appear to be history-dependent, is roughly the time scale for a random trajectory that has entered the state to *lose memory* of how it entered. This time corresponds to a relaxation or equilibration time within a state and obviously depends on the nature of the state. In general, states that include larger amounts of phase space, or that have large internal (free) energy barriers, will require longer periods of time for their internal equilibration. For a Markov description of an entire system to work, the appropriate time interval of the transition matrix must be at least as large as the *longest* equilibration time among the states that are being used to descibe the process.

Another relevant time for this type of modeling is the time for the overall system to relax to equilibrium from any arbitrary starting state. Markov models are really only interesting for times that are short relative to this time. A Markov transition matrix for times of this length or longer will take any arbitrary starting probability distribution to equilibrium in a single step. It is possible for one to have inadvertently defined states for a system such that there are some whose internal relaxation time is as long as the relaxation time of the entire system. For such systems, any Markov model can hardly be expected to be useful or accurate.

Despite these difficulties, a Markov analysis, *if it can be shown to be appropriate*, has many attractive features. First, it provides a concise way to represent information derived from many MD trajectories. Second, each of these trajectories can, in principle, be much shorter than the time for the protein to evolve from an extended state to a folded state and can be performed independently using grid, distributed or parallel computing. Third, extrapolation of the short time behavior to long times can provide information about folding rates that could be compared with experimental observations.

Fersht[43] has questioned the validity of deducing information about phenomena that occur on long time scales from large numbers of short time scale simulations. His point is that for folding simulations that are started from ensembles of starting states that are not in thermal equilibrium, there are likely to be *lag* phases during which equilibrum populations of states along various pathways are established. During this time, one may

observe anomalous pathways and kinetics, which are not representative of the dominant relaxation mechanisms. In other words, short simulations may overemphasize pathways and their associated kinetics that are important *only* during this early local equilibration phase, but that are not representative of what is observed on biologically important time scales (milliseconds or longer). However, we feel that the approach developed in this paper addresses these questions in two ways. First, we are careful to start simulations from diverse and properly Boltzmann weighted conformations. Second, fundamental to our analysis is the idea that one must examine temporal behavior on progressively longer time scales and look for convergence with respect to a number of properties before believing in an ability to predict long time behavior based on the simulations.

The structure of this paper is as follows. In section 2 we present the theory. This section has a very brief review of relevant properties of Markov chains and transition matrices, then the derivation of important correlation function, transition function and lifetime expressions. The theory section includes how these functions can be computed using biased choices for starting states, along with a reweighting scheme to restore the required Boltzmann weighting. In section 3 we provide a simple idealized example and examine it with the techniques described in section 2. The example is designed to demonstrate that a system that does not exhibit Markovian behavior on short time scales can appear Markovian on sufficiently long time scales. Section 4 is a summary of our findings and a discussion of future directions.

2. Theory

2.1. Properties of Markov Chains. Several important properties of discrete time Markov chains will be summarized here. For a comprehensive review of the subject there are numerous excellent texts available.[44,45] Consider a transition matrix, $\mathbf{T}(\tau)$, whose (i, j) element is defined as the probability that, when the system is prepared in state j at some time, it will be in state i when observed some time τ later. Consider also a vector $\mathbf{P}(t)$, whose ith element is the probability of finding the system in state i at time t. For a true Markov process, the following holds:

$$\mathbf{P}(t+\tau) = \mathbf{T}(\tau)\,\mathbf{P}(t) \tag{1}$$

For discrete times, $n\tau$, we can write this as follows:

$$\mathbf{P}((n+1)\tau) = \mathbf{T}(\tau)\,\mathbf{P}(n\tau) \tag{2}$$

Because $\mathbf{T}$ is a matrix of probabilities, its elements are nonnegative and its columns sum to unity, a property that implies that the eigenvalues, μ_i, of $\mathbf{T}$, which may in general be complex, have $|\mu_i| \leq 1$. We let Φ_i represent a right eigenvector of $\mathbf{T}$ with eigenvalue μ_i. In fact, (at least) one eigenvalue is unity, and the corresponding eigenvector is special. Without loss of generality, we will refer to these with index $i = 1$. When appropriately normalized, Φ_1 represents the steady-state distribution, because application of $\mathbf{T}$ leaves this eigenvector unchanged. For ergodic systems, where every state can eventually be reached from any other, there is only one such unit-valued eigenvalue. In this case, we can identify Φ_1 as the equilibrium distribution, $\mathbf{P}(t=\infty)$.

Detailed balance is the condition that the flux in probability between any pair of states is equal in each direction when the system is at equilibrium:

$$T_{i,j}P_j(t=\infty) = T_{j,i}P_i(t=\infty) \tag{3}$$

This condition is usually satisfied by physical systems, including ones governed by Newtonian dynamics such as ours. The only requirement to achieve this property is that the dynamics be energy conserving and time-reversible.

Detailed balance is a very strong condition and results in several consequences. First, all the eigenvalues and eigenvectors of $\mathbf{T}$ are real. Second, the eigenvectors form a complete set, and they can be used to express *any* solution to eq 2 as a linear combination of the eigenvectors as follows:

$$P(n\tau) = \sum_{i=1} c_i \mu_i^n \Phi_i \tag{4}$$

Third, the eigenvectors are orthogonal under the following definition for the inner product in a vector space whose dimension is the number of states:

$$(\phi,\psi) = \sum_i \frac{\phi_i \psi_i}{\Phi_{1,i}} \tag{5}$$

(Here, $\Phi_{1,i}$ is the ith component of Φ_1, which is the steady-state probability that the system is in state i.) With this, we can uniquely determine the coefficients c_i to generate a solution $\mathbf{P}(t=n\tau)$ that satisfies any arbitrary initial condition, $\mathbf{P}(t=0)$ as follows:

$$c_i = (\Phi_i, \mathbf{P}(t=0)) \tag{6}$$

Thus, the behavior of a probability distribution can be described as if it were a sum of *modes*, each with a different temporal behavior, related to its associated eigenvalue. A mode with an eigenvalue of $\mu < 1$ exhibits exponential decay,[46] with an exponential decay constant given by $-\tau/\ln \mu$. The mode with an eigenvalue of unity has no temporal change and, therefore, corresponds to the steady state, or stationary distribution. The other modes correspond to "probability fluxes", with varying rates of change. The mode with the largest eigenvalue less than unity is the slowest and is important because it determines the time limiting processes in the system. The smallest eigenvalues are associated with the modes that have the shortest relaxation times, because they decay quickly from any arbitrary starting distribution.

An important attribute of a Markov chain is that it is also a Markov chain when viewed on a coarser time scale. For example, consider the Markov chain produced by the transition matrix $\mathbf{T}(\tau)$. If this chain were observed on a coarser time scale, such as at a temporal resolution of $n\tau$, it would be indistinguishable from one produced by a different Markov transition matrix $\mathbf{S} = \mathbf{T}^n$. In fact, it is easy to show that the eigenvalues of $\mathbf{S}$ can be obtained by raising the eigenvalues of $\mathbf{T}$ to the nth power. (The eigenvectors of $\mathbf{T}$ and $\mathbf{S}$ are the same.)

This property of a Markov chain can be exploited to help determine whether a process is indeed Markovian. If one were to examine a chain at varying temporal resolutions, $n\tau$ for $n = 1, ...,$ and deduced a transition matrix $\mathbf{S}_n$ for each of these temporal resolutions, the eigenvalues of these matrices would be related. In fact, if $\mu_i(n)$ are the eigenvalues of transition matrix $\mathbf{S}_n$, a plot of $-n\tau/\ln \mu_i(n)$ as a function of n should show a set of constant functions. This is because at temporal resolution $n\tau$, the behavior would be described by a transition matrix $\mathbf{S} = \mathbf{T}^n$, with eigenvalues $\mu_i{}^n(n=1)$.

Another important property of Markov chains that can be exploited to test for Markovian behavior is the distribution of lifetimes observed for each of the states. This distribution follows an easily described formula. Diagonal element $T_{ii}(\tau)$ of transition matrix **T** determines the probability of seeing the system in state i at some time, given that it was in state i at a time τ earlier. The probability of *not* being in state i at some time, given that it was in state i at time τ earlier, is $1 - T_{ii}$. So, given that the system was in state i at time $t = 0$, the probability that it will be observed to remain in state i for exactly $L - 1$ more consecutive observations and then be in some other state on the Lth observation is given by $T_{ii}^{L-1}(1-T_{ii})$. Using this probability, one can show that the mean lifetime of state i is $1/(1 - T_{ii})$. States with observed lifetime distributions that are different from this are exhibiting non-Markovian behavior.

2.2. Microstates and Macrostates. We wish to relate the above discussion to molecular dynamics simulations. We will define a *microstate* to be a specification of all the coordinates and momenta of a system. For an N-particle system, there are $3N$ coordinate and $3N$ momentum components. For this discussion we will represent a microstate as x, with the understanding that this is a $6N$-component vector.

The probability of finding the system in a state where the coordinates and momenta are in a volume element dx about x is given by

$$P(x)\,\mathrm{d}x = \frac{\mathrm{e}^{-\beta H(x)}\,\mathrm{d}x}{\int \mathrm{d}x\,\mathrm{e}^{-\beta H(x)}} \tag{7}$$

where $H(x)$ is the Hamiltonian for the system, and $\beta = 1/k_{\mathrm{B}}T$. Note that this defines $P(x)$ as a probability density.

We define *macrostates* as collections of microstates that have some attribute in common. Formally, we can define a set of indicator functions, $\Omega^{(i)}(x)$, which allow us to classify microstates as to which macrostate they belong.

$$\Omega^{(i)}(x) \equiv \begin{cases} 1 & \text{if microstate } x \text{ is in macrostate } i \\ 0 & \text{if not} \end{cases} \tag{8}$$

The set of Ω functions must span all of space in a nonoverlapping way, such that every possible microstate x must belong to exactly one macrostate. This means that

$$\sum_i \Omega^{(i)}(x) = 1 \qquad \text{for any and all } x \tag{9}$$

The probability of finding the system in any microstate that is consistent with some particular macrostate i is proportional to the volume of thermally accessible phase space consistent with that macrostate.

$$P^{(i)} = \int \mathrm{d}x\, P(x)\, \Omega^{(i)}(x) \tag{10}$$

$$= \frac{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)} \Omega^{(i)}(x)}{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)}} \tag{11}$$

$$= \langle \Omega^{(i)} \rangle \tag{12}$$

We could have restricted the integral to regions of phase space that are consistent with macrostate i, and then left out the Ω. But by use of the Ω function, we can formally let the range of the integral go over all of phase space, and only microstates consistent with macrostate i are counted.

Note that the condition in eq 9 on the $\Omega^{(i)}$ functions, that each microstate be assigned to exactly one macrostate, provides a normalization condition for the $P^{(i)}$.

$$\sum_i P^{(i)} = \frac{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)} \sum_i \Omega^{(i)}(x)}{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)}} \tag{13}$$

$$= 1 \tag{14}$$

2.3. Describing Dynamical Processes. Now consider trajectory data, where the phase space point propagates in time according to, for example, Newton's equations. By virtue of the evolution of $x(t)$, $\Omega^{(i)}(x)$ is now also an implict function of time. We will be especially interested in two types of functions. The first type are the correlation functions, $C_{ij}(\tau)$, defined as follows:

$$C_{ij}(\tau) \equiv \langle \Omega^{(i)}(x(\tau))\, \Omega^{(j)}(x(0)) \rangle \tag{15}$$

$$= \frac{\int \mathrm{d}x(0)\, \mathrm{e}^{-\beta H(x(0))} \Omega^{(i)}(x(\tau))\, \Omega^{(j)}(x(0))}{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)}} \tag{16}$$

This gives the *joint* probability of finding the system in macrostate i at one time, and in macrostate j at some time τ earlier. In the limit of $\tau \rightarrow 0$ the correlation function becomes $\langle \Omega^{(i)} \rangle \delta_{ij} = P^{(i)} \delta_{ij}$. At very long times, the probabilities of being in these macrostates becomes uncorrelated, and approaches $\langle \Omega^{(i)} \rangle \langle \Omega^{(j)} \rangle = P^{(i)} P^{(j)}$. By virtue of eq 10, summing over the first and second indices gives the following relationships, true for any time, τ:

$$\sum_i C_{ij}(\tau) = \langle \Omega^{(j)}(x) \rangle \tag{17}$$

$$\sum_j C_{ij}(\tau) = \langle \Omega^{(i)}(x) \rangle \tag{18}$$

The second type of functions are transition functions, $T_{ij}(\tau)$, defined as follows:

$$T_{ij}(\tau) = \frac{\int \mathrm{d}x(0)\, \mathrm{e}^{-\beta H(x(0))} \Omega^{(i)}(x(\tau))\, \Omega^{(j)}(x(0))}{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)} \Omega^{(j)}(x)} \tag{19}$$

This gives the *conditional* probability of finding the system in macrostate i at one time, given that it was in macrostate j at some time τ earlier. In the limit of $\tau \rightarrow 0$ the transition function becomes δ_{ij}. In the long time limit, the function approaches $\langle \Omega^{(i)} \rangle = P^{(i)}$ (the limit where the probability of being in macrostate i does not depend on where the system started out). Notice that summing over the first index produces unity for any value of τ and for any value of the second index:

$$\sum_i T_{ij}(\tau) = 1 \tag{20}$$

Clearly, the correlation and transition functions are related through the following Bayes relationship:

$$T_{ij}(\tau) = C_{ij}(\tau)/P^{(j)} \tag{21}$$

In principle, one would compute C_{ij} by first selecting a very large set of starting states, $x(0)$, that are Boltzmann weighted over the entire phase space. (By this, we mean that configuration $x(0)$ occurs in the set with a probability proportional to $\exp[-\beta H(x(0))]$.) These starting states would be produced through some sort of canonical sampling method, such as Monte Carlo or canonical ensemble molecular dynamics.[47,48] Next, one would note which macrostate each of the starting states was in. One would then use energy conserving molecular dynamics to evolve these starting states for a time τ and note which macrostate they ended up in. $C_{ij}(\tau)$ is the fraction of all of these trajectories that started in macrostate j and ended up in i. However, because each of these trajectories is energy-conserving, they maintain their relative Boltzmann weights over time. Therefore, a given trajectory, as it passes through various macrostates, can be used to obtain information about many different transition functions, and for many different times, τ. For example, where each trajectory was at time t and at time $t + \tau$ could be noted and used with equivalent weight in computing the $C_{ij}(\tau)$.

Suppose we have M Boltzmann weighted starting states from which trajectories $x_m(t)$, $m = 1, ..., M$, have been computed for times from $t = 0$ to $t = T_m$. From these we can estimate $C_{ij}(\tau)$ with the following:

$$C_{ij}(\tau) \simeq \frac{1}{M}\sum_{i}^{M}\frac{1}{T_m - \tau}\int_0^{T_m-\tau} \mathrm{d}t\, \Omega^{(i)}(x_m(t+\tau))\, \Omega^{(j)}(x_m(t)) \tag{22}$$

$$= \frac{1}{M}\sum_{m=1}^{M}\overline{\Omega^{(i)}(x_m(\tau))\, \Omega^{(j)}(x_m(0))} \tag{23}$$

In this expression, note that longer trajectories do not get greater weight in the sum. To do so would upset the desired Boltzmann weighting. (The overbar notation will be used to represent time averages over some trajectory, and the trajectory over which the average is to be taken will be indicated by what is beneath the overbar.) Generally, of course, trajectories are sampled at discrete times and the time integral above is evaluated as a sum over these samples. Therefore, the $C(\tau)$ functions are actually evaluated at multiples of this sampling period.

We have many ways to evaluate $P^{(j)}$. A particulary useful one makes use of the desire to enforce the relationship in eq 17.

$$P^{(j)}(\tau) = \sum_i C_{ij}(\tau)$$

$$\simeq \frac{1}{M}\sum_{m=1}^{M}\frac{1}{T_m - \tau}\int_0^{T_m-\tau}\mathrm{d}t\,[\sum_i \Omega^{(i)}(x_m(t+\tau))]\Omega^{(j)}(x_m(t))$$

$$\simeq \frac{1}{M}\sum_{m=1}^{M}\frac{1}{T_m - \tau}\int_0^{T_m-\tau}\mathrm{d}t\, \Omega^{(j)}(x_m(t)) \tag{24}$$

(Note that this approximation for $P^{(j)}$ depends on τ.) With eqs 23 and 24 for C_{ij} and $P^{(j)}$, eq 21 is used to compute T_{ij}. When used in this context, the approximation for $P^{(j)}$ in eq 24 results in the desired normalization in eq 20 for T_{ij}. The final result for T_{ij} is as follows:

$$T_{ij}(\tau) \simeq \frac{\frac{1}{M}\sum_{m=1}^{M}\frac{1}{T_m - \tau}\int_0^{T_m-\tau}\mathrm{d}t\, \Omega^{(i)}(x_m(t+\tau))\, \Omega^{(j)}(x_m(t))}{\frac{1}{M}\sum_{m=1}^{M}\frac{1}{T_m - \tau}\int_0^{T_m-\tau}\mathrm{d}t\, \Omega^{(j)}(x_m(t))} \tag{25}$$

$$= \frac{\frac{1}{M}\sum_{m=1}^{M}\overline{\Omega^{(i)}(x_m(\tau))\, \Omega^{(j)}(x_m(0))}}{\frac{1}{M}\sum_{m=1}^{M}\overline{\Omega^{(j)}(x_m(0))}} \tag{26}$$

We will often refer to the argument of a correlation or transition function as the *lag* time, because it refers to some time period we *wait* before characterizing the system, after having seen the system in some condition earlier.

2.4. Distributions of Lifetimes. We are interested in computing the observed lifetime distributions for various states. Consider a "counting" function of x, $K_L^{(i)}(x;\tau)$, that is unity only if microstate x is in state i at times $t = 0, \tau, 2\tau, ..., (L-1)\tau$, and is *not* in state i at time $t = L\tau$. An expression for this function would be as follows:

$$K_L^{(i)}(x;\tau) = \Omega^{(i)}(x(0))\, \Omega^{(i)}(x(\tau))\, \Omega^{(i)}(x(2\tau)) ... \tag{27}$$

$$\times\, \Omega^{(i)}(x((L-1)\tau))\,(1 - \Omega^{(i)}(x(L\tau))) \tag{28}$$

This is an indicator function for microstates that are observed to be in a particular macrostate i for L consecutive observations that are spaced by τ in time. Note that with this definition, if we sum over all values of L, we get the following:

$$\sum_{L=1}^{\infty} K_L^{(i)}(x;\tau) = \Omega^{(i)}(x) \tag{29}$$

because if microstate x is in macrostate i, it will eventually leave, and one of the $K_L^{(i)}$ functions will evalute to unity, and if microstate x is not in macrostate i, none of them will, so that $K_L^{(i)}$ will evaluate to zero for every value of L.

Using $K_L^{(i)}$, an expression can be made for the thermally accessible fraction of phase space in macrostate i that survives for L consecutive observations at times $t = 0,\tau, ..., (L-1)\tau$ before leaving state i. This is given by the following:

$$\langle K_L^{(i)}(x;\tau)\rangle_i = \frac{\int \mathrm{d}x(0)\, \mathrm{e}^{-\beta H(x(0))}\Omega^{(i)}(x(0))\, K_L^{(i)}(x(0);\tau)}{\int \mathrm{d}x(0)\, \mathrm{e}^{-\beta H(x(0))}\Omega^{(i)}(x(0))} \tag{30}$$

Note that using this relationship and eq 29 it can be seen that

$$\sum_{L=1}^{\infty}\langle K_L^{(i)}(x;\tau)\rangle_i = 1 \tag{31}$$

The set of $\langle K_L^{(i)}\rangle_i$ for different values of L therefore provides a Boltzmann weighted and normalized distribution of lifetimes for microstates originating in macrostate i. The mean lifetime of microstates in macrostate i, which may also be called the mean lifetime of macrostate i is given by

$$L^{(i)} = \sum_{L=1}^{\infty} L\langle K_L^{(i)}\rangle_i \tag{32}$$

This lifetime is measured in units of τ, the observation period.

We will not evaluate $\langle K_L^{(i)}\rangle_i$ directly but, rather,

$$\langle K_L^{(i)}(x;\tau)\rangle = \frac{\int \mathrm{d}x(0)\, \mathrm{e}^{-\beta H(x(0))} K_L^{(i)}(x(0);\tau)}{\int \mathrm{d}x(0)\, \mathrm{e}^{-\beta H(x(0))}} \tag{33}$$

and then derive $\langle K_L^{(i)}\rangle_i$ from it as follows:

$$\langle K_L^{(i)}(x;\tau)\rangle_i = \langle K_L^{(i)}(x;\tau)\rangle / P^{(i)} \tag{34}$$

where $P^{(i)}$ is given by eq 12.

To evaluate these ensemble averages, we use time averages over a set of microcanonical trajectories that were started from a Boltzmann weighted set of starting states. Suppose we have M Boltzmann weighted starting states from which trajectories $x_m(t)$, $m = 1, ..., M$, have been computed for times from $t = 0$ to $t = T_m$. From these we can estimate $\langle K_L^{(i)}(x;\tau)\rangle$ with the following:[49]

$$\langle K_L^{(i)}(x;\tau)\rangle \simeq \frac{1}{M}\sum_{m=1}^{M} \frac{1}{T_m}\int_0^{T_m} \mathrm{d}t\, K_L^{(i)}(x_m;\tau) \tag{35}$$

$$= \frac{1}{M}\sum_{m=1}^{M} \overline{K_L^{(i)}(x_m;\tau)} \tag{36}$$

An important aspect of this equation is that it produces lifetime distributions that are parametrically dependent on a time interval, τ, which is related to the period between consecutive observations. We will see in an example that the qualitative nature of the lifetime distribution can change with this time interval.

2.5. Higher Order Transition Matrices and Correlation Functions. Later we will establish the degree to which transition probabilities are *history independent*. This property is a prerequisite condition if the observed transition matrices are to be used as Markov transition matrices to infer long time behavior of the system.

Consider the conditional probability, $T_{i|jk}$, of observing the system in state i at time $t = 2\tau$, given that it was in state j at time $t = \tau$ *and* in state k at time $t = 0$, expressed by the following:

$$T_{i|jk}(\tau) \equiv \frac{\int \mathrm{d}x(0)\, \mathrm{e}^{-\beta H(x(0))} \Omega^{(i)}(x(2\tau))\, \Omega^{(j)}(x(\tau))\, \Omega^{(k)}(x(0))}{\int \mathrm{d}x(0)\, \mathrm{e}^{-\beta H(x(0))} \Omega^{(j)}(x(\tau))\, \Omega^{(k)}(x(0))} \tag{37}$$

We can define corresponding three-time correlation functions as follows:

$$C_{ijk}(\tau) \equiv \frac{\int \mathrm{d}x(0)\, \mathrm{e}^{-\beta H(x(0))} \Omega^{(i)}(x(2\tau))\, \Omega^{(j)}(x(\tau))\, \Omega^{(k)}(x(0))}{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)}} \tag{38}$$

so that

$$T_{i|jk}(\tau) = \frac{C_{ijk}(\tau)}{C_{jk}(\tau)} = \frac{C_{ijk}(\tau)}{T_{jk}(\tau)P^{(k)}} \tag{39}$$

Note that in the limit $\tau \to 0$, $C_{ijk} \to \delta_{ij}\delta_{jk}P^{(i)}$, so that in this same limit, $T_{i|jk}$ approaches $\delta_{ij}\delta_{jk}$. At long times $C_{ijk} \to P^{(i)}P^{(j)}P^{(k)}$, so $T_{i|jk} \to P^{(i)}$.

For Markovian behavior to emerge, we require that the transition probability for the system to go from state j to i be independent of the state it was in earlier. This is equivalent to the requirement for $T_{i|jk}(\tau)$ to be independent of k. Clearly, this will *not* be the case for short times, because we know that for short enough times $T_{i|jk}$ has very different values if k is not equal to j than if it is. However, for long enough times we might expect systems leaving state j to behave in a history-independent way. Therefore, we would like to find the time after which the subensemble of states that are in state k at time $t = 0$ and in state j at time $t = \tau$ have the same probability of being in state i at time $t = 2\tau$, irrespective of k.

2.6. Biased Sampling by the Use of Selection Cells. In what follows we assume that a set of states have been generated that are Boltzmann distributed. This can be done by a number of methods. For example, replica exchange Monte Carlo simulations[28,48] are often performed by this research group[30] to produce a set of N_R Boltzmann distributed states. We represent these states with the notation $\{x_{R,i},\ i = 1, ..., N_R\}$, with the R to emphasize that they are members of the set that may have been produced by a prior replica exchange simulation.

Canonical ensemble, or thermal, averages of any property, $A(x)$, may be approximated by simply averaging $A(x)$ over this set of N_R states:

$$\langle A(x)\rangle \equiv \frac{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)} A(x)}{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)}} \simeq \frac{1}{N_R}\sum_{i=1}^{N_R} A(x_{R,i}) \tag{40}$$

However, some properties A may be too expensive to evaluate over the entire set. This is certainly the case for properties that are functions of time, such as correlation functions, where the relevant ensemble average would need to be evaluated by computing trajectories produced using the N_R Boltzmann-weighted configurations as starting states. Furthermore, given that we may be averaging over a subset of the available configurations, we might also be interested in focusing the sampling in phase space on regions of particular interest, and/or where the probability density is somewhat low. A prime example would be to focus sampling near regions that are suspected transition states to improve our characterization or understanding of a kinetic process.

Therefore, in this section we describe a way to focus sampling in various multiple regions of interest by sampling uniformly from various subsets of the N_R configurations, and then by applying a weighting to regain the desired canonical ensemble average.

First, define *selection cells* as such regions of interest that can be characterized in terms of some function of x. For example, one may wish to identify a region of phase space as those states that have simultaneously a particular range of values for the radius of gyration, and some range of values for the distance between a particular donor−acceptor atom pair that is capable of hydrogen bonding. As a second example, consider the region of phase space with states that have exactly four hydrogen bonds, defined with respect to some geometric criteria. We will require that each point in phase space be assignable to at least one selection cell. (One convenient way to achieve this is to define one selection cell that corresponds to all of phase space. Of course, other ways to achieve this are also possible.) A final point about selection cells is that they do not have to have any particular relationship to the macrostate definitions described above. One key difference is that selection cells describe regions of phase space that can overlap, whereas the macrostates described above cannot overlap; each phase space point must be assigned to one and only one macrostate but can lie in multiple selection cells.

Suppose N_S such selection cells have been characterized. Because selection cells are characterized on the basis of x, there is an indicator function for each:

$$\Gamma^{(i)}(x) \equiv \begin{cases} 1 & \text{if state } x \text{ is consistent with selection cell } i \\ 0 & \text{if not} \end{cases} \tag{41}$$

It will also be convenient to define a counting function $N(x)$ that indicates how many selection cells a particular phase space point belongs to

$$N(x) = \sum_{i=1}^{N_S} \Gamma^{(i)}(x) \tag{42}$$

$N(x) \geq 1$ for any state x.

Of the N_R available states, $N_R^{(i)}$ are considered to be consistent with the definition of selection cell i, where

$$N_R^{(i)} = \sum_{j=1}^{N_R} \Gamma^{(i)}(x_{R,j}) \tag{43}$$

Note that because the regions of phase space corresponding to the various selection cells might be overlapping, and because every state is assignable to at least one selection state, the $N_R^{(i)}$ sum to a number greater than or equal to N_R.

The procedure continues by *randomly* selecting (with replacement) for each selection cell i some number, $N_S^{(i)}$, of states from among the $N_R^{(i)}$ available to the selection cell. Denote these states $\{x_{R,j}^{(i)}, i = 1, ..., N_S; j = 1, ..., N_S^{(i)}\}$. Note that the resulting sets can have some duplicated states. We can denote by M, the total number of starting states selected (counting duplicates). This is just the sum over all N_S selection cells of $N_S^{(i)}$.

The resulting sample is no longer Boltzmann weighted, even though the underlying sample was. However, with proper reweighting, Boltzmann averages can be obtained as follows:

$$\begin{aligned} \langle A(x) \rangle &= \frac{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)} A(x) \left[\sum_i \Gamma^{(i)}(x) / \sum_i \Gamma^{(i)}(x) \right]}{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)}} \\ &= \frac{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)} (A(x)/N(x)) \sum_i \Gamma^{(i)}(x)}{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)}} \\ &= \sum_{i=1}^{N_S} \frac{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)} (A(x)/N(x))\, \Gamma^{(i)}(x)}{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)} \Gamma^{(i)}(x)} \frac{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)} \Gamma^{(i)}(x)}{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)}} \\ &= \sum_{i=1}^{N_S} \langle A(x)/N(x) \rangle_i P_S^{(i)} \\ &\simeq \sum_{i=1}^{N_S} \frac{1}{N_S^{(i)}} \sum_{j=1}^{N_S^{(i)}} [A(x_{R,j}^{(i)})/N(x_{R,j}^{(i)})] (N_R^{(i)}/N_R) \end{aligned} \tag{44}$$

where $\langle\ \rangle_i$ represents an ensemble average over the region of phase space consistent with selection cell i, and $P_S^{(i)}$, the probability of finding a state in selection cell i, is approximated by $N_R^{(i)}/N_R$, the fraction of replica exchange states that are consistent with selection cell i.

This expression provides the desired weighting for averages of A over the (non-Boltzmann) set of states that were selected through the selection cell approach. The net effect is that each starting state simply has a weight associated with it. Properties are computed for this starting state, or averaged over trajectories started from this starting state, and these properties are simply summed up using these weights to produce properly Boltzmann-weighted averages. Note that these formulas reduce to more familiar ones in the limit where there is only one selection cell (covering the entire phase space), or when the selection cells do not overlap.

2.7. Computing Correlation and Transition Functions and Lifetime Distributions. We can apply the results in eq 44 of the previous section to the computation of the correlation and transition functions. We do this by associating the function $A(x)$ with the indicator functions, $\Omega^{(i)}$, and with their correlation, $\Omega^{(i)}(x(\tau))\ \Omega^{(j)}(x(0))$. For the correlation functions, eq 23 becomes

$$C_{ij}(\tau) \simeq \sum_{l=1}^{N_S} \frac{1}{N_S^{(l)}} \sum_{m=1}^{N_S^{(l)}} \frac{N_R^{(l)}}{N_R N(x_{R,m}^{(l)})} \overline{\Omega^{(i)}(x_{R,m}^{(l)}(\tau))\ \Omega^{(j)}(x_{R,m}^{(l)}(0))} \tag{45}$$

For the purposes of using it to compute the transition functions, we approximate $P^{(j)}$. After the proper weighting is applied, eq 24 becomes

$$P^{(j)}(\tau) \simeq \sum_{l=1}^{N_S} \frac{1}{N_S^{(l)}} \sum_{m=1}^{N_S^{(l)}} \frac{N_R^{(l)}}{N_R N(x_{R,m}^{(l)})} \overline{\Omega^{(j)}(x_{R,m}^{(l)}(0))} \tag{46}$$

(Recall that the time average of $P^{(i)}$ in this expression depends on τ.) The transition functions are obtained as ratios of these functions through eq 21.

We can also use the results of the previous section to compute the lifetime distributions for each state. First, the ensemble

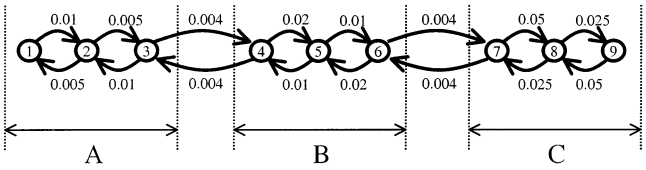

Figure 1. Representation of the Markov transition matrix for the example problem described in the text. Only "off-diagonal" values of the matrix are explicitly shown. The diagonal elements can be deduced from the fact that the sum of all transition probabilities out of any state is unity. The nine *microstates* of the nine-state system are indicated by numbers. Three *macrostates*, indicated by letters, are formed by lumping groups of three microstates together, as shown.

average of the counting function $K_L^{(i)}$ is approximated by the following:

$$\langle K_L^{(i)}(x;\tau)\rangle \simeq \sum_{l=1}^{N_S} \frac{1}{N_S^{(l)}} \sum_{m=1}^{N_S^{(l)}} \frac{N_R^{(l)}}{N_R N(x_{R,m}^{(l)})} \overline{K_L^{(i)}(x_{R,m}^{(l)};\tau)} \qquad (47)$$

To get the normalized lifetime distribution for macrostate *i*, this expression is divided by $P^{(i)}$ as given in eq 46.

3. Example

In this section, we will study a simple example designed to illustrate a number of points raised in the Introduction. This example will include an analysis as suggested in the previous section. The example system consists of a set of states with dynamics controlled by an associated Markov transition matrix. If a chain of states is generated by use of the matrix, the resulting trajectory can be recognized as Markovian. Through the process of lumping sets of states together, the Markovian nature of the process is lost on short time scales, and this loss is recognized from observations suggested in the preceding section. On longer time scales, however, the Markovian nature of the process is regained.

Consider a system with nine "microstates" and a nine-by-nine Markov transition matrix, as illustrated in Figure 1. The probability of being in each of these states can be described by a nine vector, and the discrete time evolution of this probability density can be generated by repeated multiplication of the matrix and the vector. We can also use the transition matrix to generate *trajectories* of the Markov process in state space. Using different random number seeds and starting states, one can easily generate many such trajectories. In this example, we have generated 100 trajectories of 10 000 states. These trajectories can be subjected to the analysis presented in the previous section. For example, transition functions for lag times from $t = 0$ to $t = 200$ sampling periods can be computed from the trajectories, and one set is shown in Figure 2. The curves in this figure show the probability of observing the system to be in various states as a function of time, given that it was in state 5 at time $t = 0$. For a nine-state system there are 81 such transition functions. By taking one element, corresponding to a particular time, from each function, one can construct a transition matrix that describes the evolution of the system over that period of time. Because we have evaluated the functions at 201 times, we can construct 201 such transition matrices. We can, in fact, index these matrices by their lag times. (The matrix for $t = 0$ is the identity matrix.) We expect that the matrix corresponding to an evolution by one time period to strongly resemble the one used to contruct the trajectory in the first place. They are not identical, of course, due to the fact that we based our estimates of the transition functions on a finite number of finite length trajectories.

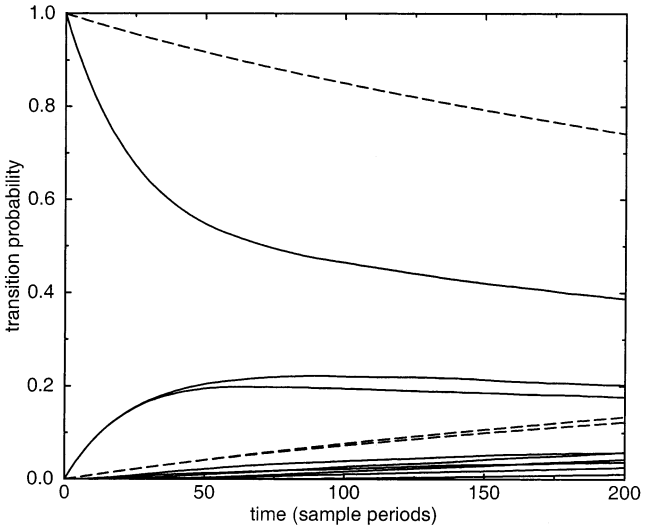

Figure 2. Transition functions, showing the probability of being in various states as a function of time given that the system was in state 5 of the nine-state system at time $t = 0$ (solid lines), or given that it was in state 2 of the lumped three-state system at time $t = 0$ (dashed lines). For the nine-state case, the rapidly decreasing function represents the diagonal function $T_{5,5}$, and the two most rapidly rising functions represent the probability of being in states 4 or 6. For the three-state system, the decreasing function represents the diagonal function $T_{B,B}$, and the two more rapidly rising functions represent the probability of being in states A or C.

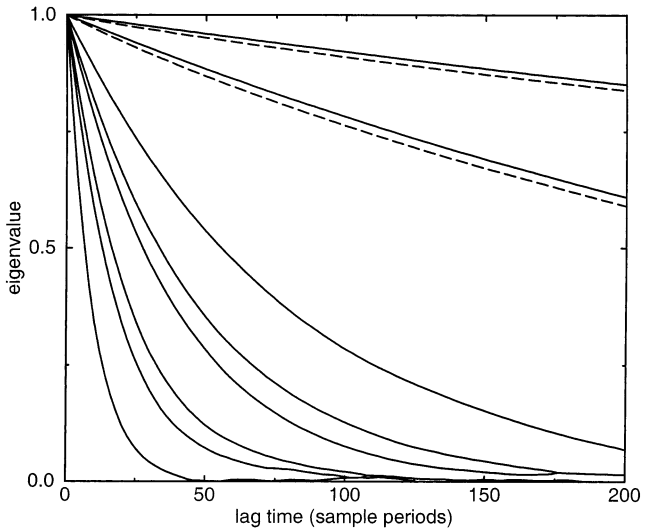

Figure 3. Eigenvalues as a function of time for a nine- (solid lines) and a three-state (dashed lines) system. Each system has one eigenvalue of unity for all lag times. The eigenvalues of the nine-state system show the expected exponential decay of a Markov process. Those of the three-state system need not, because they represent a process that is not necessarily Markovian.

The transition matrices can be diagonalized and their nine eigenvalues can be plotted as a function of these lag times. The result is shown in Figure 3. For a Markov process, we expect that the transition matrix corresponding to evolution by *n* steps to be just the transition matrix corresponding to one step raised to the power of *n*. The eigenvalues of these matrices should be similarly related, so if μ is an eigenvalue corresponding to the matrix $\mathbf{T}(t=1)$, we expect there to be an eigenvalue with value μ^n for the matrix corresponding to $\mathbf{T}(t = n)$. The figure shows such expected exponential decay, characteristic of a Markov process. The eigenvalues describe the rates of decay of various relaxation processes in the system. An eigenvalue of μ associated with a matrix that corresponds to evolution by a lag time *t*, for

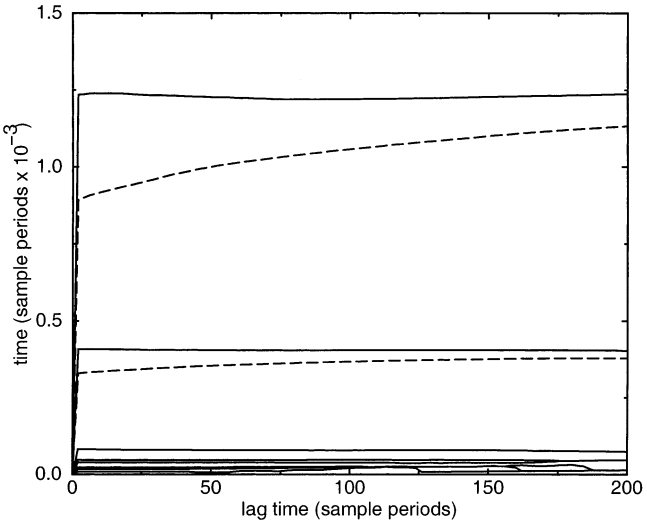

Figure 4. Time scales for relaxation processes implied by the eigenvalues of transition matrices. The exponential decay constants for the processes are related to the eigenvalues, μ, by $\tau = -t/\ln \mu$. Note that for the nine-state system (solid lines), the functions are constants of the lag time. For the lumped three-state system (dashed lines), the functions are not constant of the lag time but approach them as the lag time increases.

example, corresponds to an exponential decay time given by $\tau = -t/\ln \mu$. These decay times, as a function of the lag times of the matrices from which they came, are shown in Figure 4. Note that they are constant functions for a Markov process. The largest eigenvalues determine rates of the slowest processes in the system.

Not surprisingly, indications based on lifetime distributions (not shown) also suggest Markovian behavior for the nine-state system. The lifetime of each state is simply related to the value of a diagonal element of the transition matrix. Looking at the lifetime distributions with different lag times shows Markovian behavior.

Consider now what happens if sets of three microstates are lumped together to form macrostates. The *same* trajectories described above, produced using the full nine-state transition matrix, were analyzed as if the system were actually a three-state system. That is, when any of microstates 1−3 are observed, the system is assigned to macrostate A, etc. The correspondence between (numbered) microstates and (lettered) macrostates is shown in Figure 1. This situation mimics the fact that the underlying process of classical dynamics is Markovian when viewed in the context of infinitesimally small elements of phase space. However, in our analysis of peptide behavior, we are interested in macrostate definitions that ignore all momenta and all degrees of freedom of the solvent, and that encompass significant volumes of configuration space, such as all regions where the radius of gyration of the peptide is within some finite range of values. The macrostates we are ultimately interested in for characterizing folding dynamics, therefore, represent a significant amount of lumping. With lumping, we expect to observe non-Markovian behavior in transitions from state A to B, for example, because the probability of making a transition to state B really depends on whether the system is in the state 1, 2, or 3 compartment of A. The resulting behavior of A should be more complex than if a single transition probability described its transitions to B.

Having projected from nine down to three states, we now subject the trajectories to the same kind of analysis described in the previous section. We generate correlation functions, macrostate probabilities, transition functions, and lifetime distributions. Example transition functions are shown in Figure 2. They represent much slower processes than observed for the underlying nine-state system because they are describing transitions between more weakly coupled *manifolds* of states. The eigenvalue spectrum for the three-state system is shown in Figure 3. Notice that for this example, the larger eigenvalues of the three-state system are similar to those of the nine-state system. The times implied by the eigenvalues are shown in Figure 4. For the three-state system, these functions are not constant, a signature of non-Markovian behavior. However, as the lag time increases, the two curves for the three-state system appear to approach corresponding constant curves for the nine-state system. That is, on sufficiently long (lag) time scales, we see that the three-state system can appear to behave in a Markovian manner. Furthermore, the slow processes and their time scales in the nine-state system are being adequately described on sufficiently long time scales in the analysis of the three-state system.

Figure 5 shows lifetime distributions for two states of the three-state system. Shown on the plots are both the observed (Boltzmann-weighted) lifetime distribution and, for comparison, what would be expected if the distribution were that of a Markov process with the same mean lifetime. State A, consisting of states 1−3 of the nine-state system, exhibits non-Markovian behavior on short time scales, but as the time lag increases, we see that a Markovian description could be adequate. State B, consisting of states 4−6 of the nine-state system, appears to behave in a way consistent with a Markov process on all the time scales shown.

The nine-state transition matrix used in this example was contrived to illustrate the transition from Markovian to non-Markovian behavior upon lumping, and the transition from non-Markovian to Markovian which appears again on sufficiently long time scales. The matrix was constructed so that states 1−3 make frequent transitions among themselves, each with a mean lifetime of about 100 steps, similarly, for states 4−6, each with a mean lifetime of about 50 steps, and for states 7−9, each with a mean lifetime of about 20 steps. Transitions between the A, B, and C macrostates of the three-state system are determined largely by the small transition probabilitites between states 3 and 4, and between states 6 and 7. These determine the much longer lifetimes of the three-state system. The approach to Markovian behavior should occur on time scales that are related to both the relaxation times among the states within a macrostate and the transition times between macrostates. To see correct long time scale behavior, we need to formulate macrostates that have internal equilibration time scales that are short compared to the lifetimes of the macrostates themselves.

4. Conclusions and Discussion

We have presented a rigorous derivation of formulas for the computation of transition probabilities from molecular dynamics data. The formulation uses Boltzmann weighted conformations as starting states for microcanonical simulations. It takes into account the need for enhanced sampling around parts of phase space that might be involved in transition states through the use of a reweighting scheme that *restores* the Boltzmann weighting. We feel that it is important to start trajectories from many starting states and that these starting states should be representative of some thermodynamically meaningful ensemble, that is, they should be Boltzmann distributed, so that statistically meaningful and unbiased conclusions can be drawn about the number, nature, and relative importance of folding pathways. This is very difficult to do from studies that start from artificially

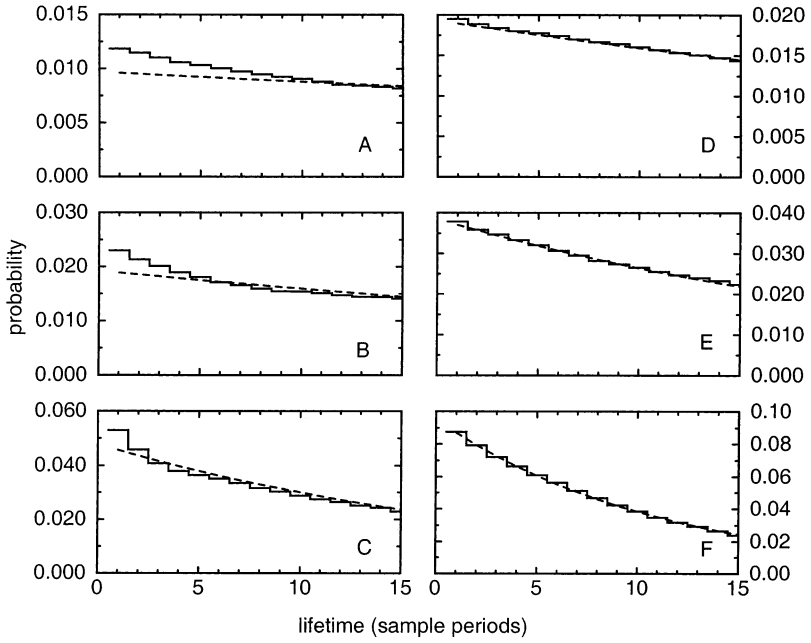

Figure 5. Lifetime distributions for the lumped three-state system. On each plot the solid line represents observed lifetime distributions and the dashed line represents the distribution that would be expected if the behavior were Markovian and had a mean lifetime that was equal to the observed mean lifetime. Panels A−C show data for state A of the three-state system using time lags of 10, 20, and 50 sampling periods, respectively. Similar data for state B of the three-state system are shown in panels D−F (again, with time lags of 10, 20, and 50).

prepared starting conformations, such as fully extended states. We also feel that from these Boltzmann distributed starting states microcanonical and time reversible (NVE) molecular dynamics simulations should be performed because the thermal control mechanisms in use for canonical sampling interfere with the dynamics of the system. It is possible that for some peptide systems, if not the majority, the kinds of thermal control in common use will have minimal or no impact on protein folding kinetic studies. This remains to be seen. But there is no scientific reason to impose thermal control in studies that use energy conserving algorithms to generate dynamics.

An important aspect of the formulation is that no prior assumption of Markovian behavior is assumed and so the degree to which the observations are Markovian can be assessed in an unbiased way. We also provide in this formulation for the possibility that observations may not be Markovian on short time scales but may be on longer time scales. Furthermore, the formulation provides a way to compute correlation and transition functions in a way that satisfies many of the desireable normalization conditions. The equilibrium distribution produced from replica exchange simulations can be used in such a way that there is a high degree of consistency between this distribution and the eigenvector of the transition matrix that corresponds to the steady-state distribution.

We have applied the techniques described to a sample problem that demonstrates how appropriately defined macrostates might behave under this kind of analysis. We would not, for example, expect to observe Markovian behavior on short time scales, but the formalism provides suggestions for metrics that might be exploited to assess the degree to which a Markovian description might be appropriate at longer time scales. These include an analysis of the eigenvalue spectrum as a function of lag time, lifetime distributions and history dependence of transition probabilities.

Work is ongoing to address the issue of better macrostate definitions, such as the formulation of an automated process for order parameter selection and binning. We also wish to address issues related to the sensitivity of our results with respect to the number and length of the dynamical simulations.

Appropriately applied, this approach has the potential to properly elucidate the behavior of protein folding from multiple independent trajectories. This requires appropriate Boltzmann weighted coverage of phase space as well as high quality energy conserving trajectories. We are looking forward to the application of these techniques to a variety of peptide and small protein systems.

In a companion paper, we describe an application of this formalism to the folding of the β-hairpin from protein G using a novel macrostate space definition that resolves not only the number, but the pattern of native hydrogen bonds.

Acknowledgment. The authors wish to acknowledge the many and very helpful discussions with Bruce Berne (Columbia University), Hans Andersen, Persi Diaconis and Vijay Pande (Stanford University), and Ken Dill and John Chodera (University of California at San Francisco).

References and Notes

(1) Swope, W. C.; Pitera, J. W.; Suits, F.; Pitman, M. C.; Eleftheriou, M.; Fitch, B. G.; Germain, R. S.; Rayshubski, A.; Ward, T. J. C.; Zhestkov, Y.; Zhou, R. *J. Phys. Chem B* **2004**, *108*, 6582.

(2) Capaldi, A. P.; Radford, S. E. *Curr. Opin. Struct. Biol.* **1998**, *8*, 86.

(3) Eaton, W. A.; Munoz, V.; Thompson, P. A.; Henry, E. R.; Hofrichter, J. *Acc. Chem. Res.* **1998**, *31*, 745.

(4) Lapidus, L. J.; Eaton, W. A.; Hofrichter, J. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 7220.

(5) Martin Gruebele, M. In *Mechanisms of Protein Folding*, 2nd ed.; Pain, R. H., Ed.; Oxford University Press: New York, 2000.

(6) Roder, H.; Elove, G. A.; Shastry, M. C. R. In *Mechanisms of Protein Folding*, 2nd ed.; Pain, R. H., Ed.; Oxford University Press: New York, 2000.

(7) Duan, Y.; Kollman, P. *Science* **1998**, *282*, 740.

(8) Snow, C. D.; Nguyen, H.; Pande, V. S.; Gruebele, M. *Nature* **2002**, *420*, 102.

(9) Snow, C. D.; Zagrovic, B.; Pande, V. S. *J. Am. Chem. Soc.* **2002**, *124*, 14548.

(10) Daura, X.; Juan, B.; Seebach, D.; v. Gunsteren, W. F.; Mark, A. E. *J. Mol. Biol.* **1998**, *280*, 925.
(11) Sheinerman, F.; Brooks, C., III. *J. Mol. Biol.* **1998**, *278*, 439.
(12) Pitera, J. W.; Swope, W. C. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 7587.
(13) Yang, W. Y.; Pitera, J. W.; Swope, W. C.; Gruebele, M. *Nature Struct. Biol.*, submitted.
(14) Mayor, U.; Johnson, C. M.; Daggett, V.; Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 13518.
(15) Dobson, C. M.; Karplus, M. *Curr. Opin. Struct. Biol.* **1999**, *9*, 92.
(16) Dinner, A. R.; Sali, A.; Smith, L. J.; Dobson, C. M.; Karplus, M. *Trends Biochem. Sci.* **2000**, *25*, 331.
(17) Mayor, M.; Guydosh, N. R.; Johnson, C. M.; Grossmann, J. G.; Sato, S.; Jas, G. S.; Freund, S. M. V.; Alonso, D. O. V.; Daggett, V.; Fersht, A. R. *Nature* **2003**, *421*, 863.
(18) Phillips, J.; Zheng, G.; Kumar, S.; Kale, L. *Supercomputing 2002 Proceedings* **2002**, http://www.sc2002.org/paperpdfs/pap.pap277.pdf.
(19) Makino, J.; Taiji, M. *Scientific simulations with special-purpose computers*; John Wiley and Sons: Chichester, U.K., 1998.
(20) Shirts, M.; Pande, V. S. *Science* **2000**, *290*, 1903.
(21) Shirts, M. R.; Pande, V. S. *Phys. Rev. Lett.* **2001**, *86*, 4983.
(22) Folding@Home home page. http://folding.stanford.edu.
(23) Globus Alliance home page. http://www.globus.org.
(24) BlueGene project home page. http://www.research.ibm.com/bluegene/.
(25) Allen, F.; et al. *IBM Syst. J.* **2001**, *40*, 310.
(26) Fitch, B. G.; Germain, R. S.; Mendell, M.; Pitera, J. W.; Pitman, M. C.; Rayshubski, A.; Sham, Y.; Suits, F.; Swope, W. C.; Ward, T. J. C.; Zhestkov, Y.; Zhou, R. *J. Parallel Distributed Computing* **2003**, *63*, 759.
(27) Adiga, N.; et al. *Supercomputing 2002 Proceedings* **2002**, http://www.sc-2002.org/paperpdfs/pap.pap207.pdf.
(28) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60*, 96.
(29) Garcia, A. E.; Sanbonmatsu, K. Y. *Proteins* **2002**, *42*, 345.
(30) Zhou, R.; Berne, B. J.; Germain, R. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 14931.
(31) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *J. Chem. Phys.* **2003**, *119*, 5740.
(32) Ozkan, S. B.; Bahar, I.; Dill, K. A. *Nature Struct. Biol.* **2001**, *8*, 765.
(33) Ozkan, S. B.; Dill, K. A.; Bahar, I. *Biopolymers* **2003**, *68*, 35.
(34) Zhang, W.; Chen, S. *J. Chem. Phys.* **2003**, *118*, 3413.
(35) Levy, Y.; Jortner, J.; Berry, R. S. *Phys. Chem. Chem. Phys.* **2002**, *4*, 5052.
(36) Hummer, G.; Kevrekidis, I. G. *J. Chem. Phys.* **2003**, *118*, 10762.
(37) de Groot, B. L.; Daura, X.; Mark, A. E.; Grubmuller, H. *J. Mol. Biol.* **2001**, *309*, 299.
(38) Becker, O. M.; Karplus, M *J. Chem. Phys.* **1997**, *106*, 1495.
(39) Buchholz, P. *J. Appl. Probability* **1994**, *31*, 59.
(40) Jernigan, R. W.; Baran, R. H. *Statistics Probability Lett.* **2003**, *64*, 17.
(41) Huisinga, W.; Schutte, C.; Stuart, A. M. *Cummun Pure Appl. Math.* **2003**, *56*, 234.
(42) Fersht, A. R. *Structure and Mechanism in Protein Science*; W. H. Freeman: New York, 1999.
(43) Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 14122.
(44) Oppenheim, I.; Shuler, K. E.; Weiss, G, H. *Stochastic Processes In Chemical Physics: The Master Equation*; The MIT Press: Cambridge MA, 1977.
(45) Van Kampen, N. G. *Stochastic Processes in Physics and Chemistry*; North-Holland Physics Publishing: New York, 1987.
(46) One can see this by noting that for temporal evolution by a time $t = n\tau$, the decay factor $\mu^{t/\tau}$ corresponds to an exponential decay factor of $\exp(-t/\tau_{decay})$. So, $\tau_{decay} = -\tau/\ln(\mu)$.
(47) Andersen, H. C. *J. Chem. Phys.* **1980**, *72*, 2384.
(48) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140.
(49) In evaluating this expression we adopt a convention where we assume that if the trajectory had gone one time step longer, it would have left the state it was in at the end of the simulation. Similarly, we assume that the state immediately before the first time step produced during the simulation is in a different state than that of the first step of the simulation.