Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 2. Example Applications to Alanine Dipeptide and a β-Hairpin Peptide†

William C. Swope* and Jed W. Pitera

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

Frank Suits, Mike Pitman, Maria Eleftheriou, Blake G. Fitch, Robert S. Germain, Aleksandr Rayshubski, T. J. C. Ward, Yuriy Zhestkov, and Ruhong Zhou

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

Received: November 10, 2003; In Final Form: March 1, 2004

In this work we demonstrate the use of a rigorous formalism for the extraction of state-to-state transition functions as a way to study the kinetics of protein folding in the context of a Markov chain. The approach is illustrated by its application to two different systems: a blocked alanine dipeptide in a vacuum and the C-terminal β-hairpin motif from protein G in water. The first system displays some of the desired features of the approach, whereas the second illustrates some of the challenges that must be overcome to apply the method to more complex biomolecular systems. For both example systems, Boltzmann weighted conformations produced by a replica exchange Monte Carlo procedure were used as starting states for kinetic trajectories. The alanine dipeptide displays Markovian behavior in a state space defined with respect to $\phi-\psi$ torsion angles. In contrast, Markovian behavior was not observed for the β-hairpin in a state space where all possible native hydrogen bonding patterns were resolved. This may be due to our choice of state definitions or sampling limitations. Furthermore, the use of different criteria for hydrogen bonding results in the apparent observation of different mechanisms from the same underlying data: one set of criteria indicate a zipping type of process, but another indicates more of a collapse followed by almost simultaneous formation of a large number of contacts. Analysis of long-lived states observed during the simulations of the β-hairpin suggests that important aspects of the folding process that are not captured by order parameters in common use include the formation of non-native hydrogen bonds and the degree and nature of salt bridge formation.

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 apply an approach that was introduced in a companion paper[1] for characterizing some aspects of protein folding kinetics to two example systems: an alanine dipeptide and the folding of a small peptide, the C-terminal β-hairpin motif from protein G.

The alanine dipeptide is of interest because it is an example of a simple biomolecular system that exhibits multiple stable conformational states. It provides a clear example of the method we are proposing for modeling conformational kinetics.

Many believe that the first step to understanding the folding of complex proteins is to fully characterize the folding of their smallest structures, helices, and sheets. The hairpin motif, a component of a β-sheet, is one of the simplest elements of protein structure. A particularly well-studied version of this is the β-hairpin motif from protein G, which has become known as the "hydrogen atom" of protein folding. It has been extensively studied experimentally[2−8] and by a variety of theoretical and computational models.[9−20] There are still open issues about the exact folding pathway and mechanism of this peptide. For example, do native hydrogen bonds form simultaneously with, before, or after the formation of a hydrophobic core made up of the side chains of four residues, two from each strand? Do helical structures play any role as precursors in the folding process? Different simulation methods and force fields have yielded different results. A recent paper[21] has proposed an additional mechanism that involves the formation and breaking of non-native hydrogen bonds through a reptation type of motion.

Most simulation studies have addressed the thermodynamics and pathways of folding rather than the kinetics of the folding process. Notable exceptions include the work of Snow,[22] which addresses the issue with large numbers of independent and short simulations performed on a distributed computing platform, and more recent work by Bolhuis,[23] where transition path sampling techniques were applied and folding rates very close to experimentally observed ones were computed.

A number of trends are increasing the amount of computational resource available for protein folding simulations. These include improved software that can efficiently exploit parallelism,[24] special purpose hardware to support biomolecular simulation,[25] and the development of new computational approaches that can exploit parallelism across distributed computational resources.[26−29] The IBM BlueGene project,[30−33] to build a massively parallel computer to investigate biomolecular processes such as protein folding, is expected to systematically

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

* To whom correspondence should be addressed.

10.1021/jp037422q CCC: $27.50
Published on Web 04/21/2004

study a variety of peptide and small protein systems. All of these trends will result in the production of large numbers of peptide trajectories. Obtaining large numbers of independent trajectories is not only a very effective way to use parallel computing technologies, it is required for statistically meaningful and reproducible results. 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.

We have developed an approach[1] that we feel can support the interpretation of molecular dynamics trajectories toward the understanding of peptide folding thermodynamics and kinetics in the context of Markov modeling. The point of the approach is to produce *transition functions* based on observations of the trajectories. From these transition functions one can construct a set of transition matrices whose properties can be examined in a way to determine whether they are appropriate to be used in a Markov description of the process. This approach has been described in a companion paper.[1] In this paper we will apply the methodology to describe the kinetics of two peptide systems as an illustration of its use, potential effectiveness, and possible limitations that must be overcome.

For the approach to work, one needs to define an appropriate state space, and this can be a major challenge.[1] Despite the 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 fully folded state, and can be performed independently using grid, distributed or parallel computing. And, third, extrapolation of the short time behavior to long times can provide information about folding rates and mechanisms that can be compared with experimental observations. There are certainly issues[34] regarding such extrapolations of long time behavior from many short simulations, and these have been discussed in detail in the companion paper.[1]

The structure of this paper is as follows. In section 2 we very briefly summarize key formulas that were derived in the companion paper. In Section 3, we describe the alanine dipeptide molecular system and the application of our method to the study its kinetics. In section 4, we describe the β-hairpin molecular system and the application of our method to the study of its kinetics. Section 5 is a summary of our findings and a discussion of future directions.

2. Theory

A *microstate* is a specification of 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.

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.

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

Of fundamental interest in this work is the computation of transition matrices that describe the temporal evolution of the system. The transition matrices are computed from transition functions and time correlation functions of the indicator functions. Formally, both of these types of functions are averages over canonical ensembles of information that is derived from energy conserving (microcanonical) trajectories.

Suppose we have M Boltzmann weighted starting states from which microcanonical trajectories $x_m(t)$, $m = 1, ..., M$, have been computed for times from $t = 0$ to $t = T_m$. From these we can estimate the time correlation function between indicator functions i and j, $C_{ij}(\tau)$:

$$C_{ij}(\tau) \equiv \langle \Omega^{(i)}(x(\tau))\, \Omega^{(j)}(x(0)) \rangle \tag{2}$$

$$= \frac{\int \mathrm{d}x(0)\, \mathrm{e}^{-\beta H(x(0))} \Omega^{(i)}(x(\tau))\, \Omega^{(j)}(x(0))}{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)}} \tag{3}$$

$$\simeq \frac{1}{M} \sum_{m=1}^{M} \frac{1}{T_m - \tau} \int_0^{T_m - \tau} \mathrm{d}t\, \Omega^{(i)}(x_m(t+\tau))\, \Omega^{(j)}(x_m(t)) \tag{4}$$

where $H(x)$ is the Hamiltonian, $\beta = 1/kT$, k is the Boltzmann constant and T is the temperature. Similarly, we can compute the probability of finding the system in a microstate that is consistent with some particular macrostate i:

$$P^{(i)} = \frac{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)} \Omega^{(i)}(x)}{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)}} \tag{5}$$

$$\simeq \frac{1}{M} \sum_{m=1}^{M} \frac{1}{T_m - \tau} \int_0^{T_m - \tau} \mathrm{d}t\, \Omega^{(j)}(x_m(t)) \tag{6}$$

The transition functions, $T_{ij}(\tau)$, are defined and computed as follows:

$$T_{ij}(\tau) \equiv \frac{\int \mathrm{d}x(0)\, \mathrm{e}^{-\beta H(x(0))}\, \Omega^{(i)}(x(\tau))\, \Omega^{(j)}(x(0))}{\int \mathrm{d}x\, \mathrm{e}^{-\beta H(x)}\, \Omega^{(j)}(x)} \tag{7}$$

$$= C_{ij}(\tau)/P^{(j)} \tag{8}$$

Transition functions give 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. 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 to be in some condition at time zero.

We will also be 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 $t = L\tau$. Using the Boltzmann weighted starting states described above we estimate $\langle K_L^{(i)}(x;\tau) \rangle$ with the following:[35]

$$\langle K_L^{(i)}(x;\tau) \rangle = \langle \Omega^{(i)}(x(0))\, \Omega^{(i)}(x(\tau))\, \Omega^{(i)}(x(2\tau)) \ldots \tag{9}$$

$$\times\, \Omega^{(i)}(x((L-1)\tau))\, (1 - \Omega^{(i)}(x(L\tau))) \rangle \tag{10}$$

$$\simeq \frac{1}{M} \sum_{m=1}^{M} \frac{1}{T_m} \int_0^{T_m} \mathrm{d}t\, K_L^{(i)}(x_m(t);\tau) \tag{11}$$

The thermally accessible fraction of phase space in macrostate i that survives for L consecutive occurrences at times $t = 0, \tau,$

..., $(L-1)\tau$ before leaving state i is given by the following:

$$\langle K_L^{(i)}(x;\tau)\rangle_i = \frac{\int \mathrm{d}x(0)\, \mathrm{e}^{-\beta H(x(0))}\, \Omega^{(i)}(x(0))\, K_L^{(i)}(x(0);\tau)}{\int \mathrm{d}x(0)\, \mathrm{e}^{-\beta H(x(0))} \Omega^{(i)}(x(0))} \tag{12}$$

$$= \langle K_L^{(i)}(x;\tau)\rangle / P^{(i)} \tag{13}$$

The significance of this is that the set of $\langle K_L^{(i)}\rangle_i$ for different values of L provides a normalized distribution of lifetimes for microstates originating in macrostate i. The mean lifetime of microstates in macrostate i is given by

$$L^{(i)} = \sum_{L=1}^{\infty} L \langle K_L^{(i)}\rangle_i \tag{14}$$

This lifetime is measured in units of τ.

An important aspect of these equations is that they produce lifetime distributions that are parametrically dependent on a time interval, τ, which is related to the period between consecutive observations.

The key results of this section are the expressions for the Boltzmann weighted macrostate probabilities, $P^{(i)}$ (eq 6), correlation functions, $C_{ij}(\tau)$ (eq 4), transition functions, $T_{ij}(\tau)$ (eq 8), and the lifetime distributions, $\langle K_L^{(i)}\rangle_i$ (eq 13), evaluated from eqs 11 and 6.

The Boltzmann weighting is actually facilitated by use of a scheme referred to in the companion paper[1] as the *selection cell* method. Because regions of phase space that are important to kinetic processes, such as those near transition states, might be very rarely observed even in long canonical molecular dynamics or Monte Carlo simulations, it is important to be able to enhance our sampling in these regions by selecting more starting states from them for microcanonical simulations. This allows for the more precise computation of transition functions that describe evolution in to and out of macrostates near these regions of phase space. If the starting states are chosen from a set that represents a Boltzmann distribution, the bias introduced by enhanced sampling near putative transition states would upset the Boltzmann weighting. The selection cell method corrects for this bias and still allows us to improve precision in transition functions associated with rarely sampled regions of phase space.

3. Alanine Dipeptide

3.1. Simulation Methods. The blocked alanine dipeptide (ACE-ALA-NME) in vacuo was simulated using the AMBER 6.0[36] simulation package with the parm96 parameter set.[37] A nonbonded cutoff was not used. SHAKE[38] was used to constrain all bonds to their equilibrium lengths with a tolerance of 10^{-5} Å. Center-of-mass momentum was removed any time velocities were reassigned.

The simulations were carried out in two phases. In the first phase, replica-exchange molecular dynamics,[39,40] simulations at nine temperatures (evenly spaced from 300 to 700 K) were carried out for a total of 100 ns per replica with exchange attempts every 10 ps. Acceptance ratios for exchange moves ranged from 68 to 86%. Velocities were randomly reassigned from a Maxwell−Boltzmann distribution at the appropriate temperature every 2 ps.[41] At each temperature, conformations were saved every 10 ps, yielding 10 000 starting states for subsequent kinetic simulations. At 500 K, these 10 000 starting states populate all five of the macrostates used for the subsequent kinetic analysis. ϕ, ψ, and θ[42] torsion angles were calculated for each saved conformation.

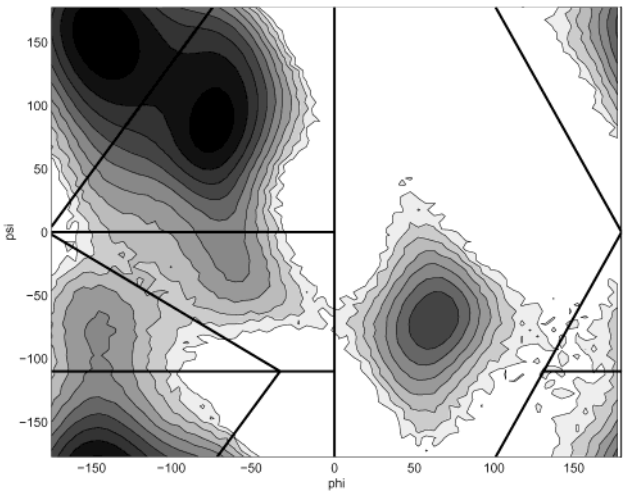

Figure 1. Free energy surface for the alanine dipeptide system in a vacuum. Lines represent constant energy contours of kT at a temperature of 500 K. The straight lines represent the boundaries of the macrostates used in this study. Horizontal boundary lines are at $\psi = 0°$ and $\psi = -110°$. The vertical boundary line is at $\phi = 0°$. Boundary vertexes are at ($\phi = \pm 180°$, $\psi = 0°$), ($\phi = 100°$, $\psi = \pm 180°$), ($\phi = -75°$, $\psi = \pm 180°$), ($\phi = -34.1666°$, $\psi = -110°$), and ($\phi = 131.1111°$, $\psi = -110°$). The corners of the graph are in macrostate 1, the basin near ($\phi = -75°$,$\psi = 90°$) is in state 2, ($\phi = -150°$, $\psi = -75°$) is in state 3, ($\phi = -50°$, $\psi = -50°$) is in state 4, and the basin near ($\phi = 60°$, $\psi = -75°$) is in state 5.

The second phase of the calculation involved using the Boltzmann weighted states from the replica-exchange simulations at 500 K as starting states for the kinetic simulations. Simulations from all 10 000 starting states were performed; no attempt was made to bias the selection of starting states from the replica-exchange data. Each kinetic simulation was 100 ps in length and coordinates were saved every 0.5 ps.

Small isolated molecular systems exhibit periodic and quasi-periodic dynamics that are not usually observed in the liquid phase.[43] The effects of isolation and of various degrees of thermal coupling on the rates of conformational change have been extensively studied. However, because the primary motivation here is to illustrate a method meant to be appropriate for the study of relatively slow molecular conformational changes in solvent, where periodic motion is not expected, all atomic velocities in the molecule were reassigned from the 500 K Boltzmann distribution every picosecond to crudely approximate the presence of a solvent bath. Relative to a single isolated molecule this will clearly have a profound effect on the kinetic behavior of this molecular system. And although our kinetic results will depend to a great degree on our choice of velocity reassignment period, this approach serves well to illustrate the methodology.

3.2. Analysis and Results. Our data set therefore consists of 200 regularly spaced conformations from each of 10 000 simulations of 100 ps. ϕ, ψ, and θ torsion angles were calculated for each saved conformation. Because the conformations are Boltzmann distributed, the observed distribution of $\phi-\psi$ angles can be used to a construct free enegy surface in these parameters. The main features of this surface are two basins corresponding to conformations usually designated $C_{7\mathrm{eq}}$, and C_{ax}. This is shown in Figure 1.

Earlier work[42] on studies of the alanine dipeptide in a vacuum and in solution has suggested that the $\phi-\psi$ torsional degrees of freedom may not be sufficient for characterizing the dynamics of this molecule. That work suggests that for the system in a vacuum, a different torsional angle must be considered, at least

TABLE 1: Fractional Populations at 500 K for the Macrostates Used for the Alanine Dipeptide Conformations[a]

			Fractional Population		
state	repex	kinetics	T(0.5 ps)	T(5.0 ps)	T(25 ps)
1	0.5733	0.5742	0.5768	0.5774	0.5776
2	0.4087	0.4081	0.4096	0.4100	0.4092
3	0.0011	0.0009	0.0009	0.0009	0.0008
4	0.0010	0.0010	0.0010	0.0010	0.0010
5	0.0159	0.0158	0.0118	0.0108	0.0114

[a] The populations were produced from the replica exchange simulations (RepEx), the kinetic simulations (Kinetics), and from the eigenvector with unit eigenvalue of transition matrixes constructed with lag times of 0.5, 5.0, and 25 ps.

for transitions between the C_{7eq} and C_{ax} conformations. This angle was labeled θ, the O−C−N−C_α angle about a peptide bond. When considered with ϕ it gave a very sharply defined transition state ensemble for the C_{7eq} to C_{ax} transition region in $\theta-\phi$ space. The point was that one would not be able to accurately describe dynamics with stochastic models that considered projections of phase space onto the ϕ and ψ dimensions. However, we have not been able to confirm this point with our simulations. We have studied the $\phi-\psi$ free energy surfaces our simulations generate as a function of θ and do not see the features that work would suggest. In fact, our $\phi-\psi$ free energy surfaces appear symmetric about a θ value of 0° near the transition state region between C_{7eq} and C_{ax}. There are other important differences in our results. The locations of our free energy minima do not exactly coincide with the locations of the energy minima reported in this eariler work. Some, if not all, of these differences might be explained by the fact that our simulations were performed at 500 K, whereas those of the earlier work were at 300 K. More importantly, the earlier work employed a different parameter set for the AMBER force field, the parm94 parameter set,[44] versus our parm96 parameter set. These force fields differ considerably in the values of ϕ and ψ torsional energy parameters and could well explain the differences we see. Consequently, we have chosen to partition our state space using the $\phi-\psi$ torsional parameters.

On the basis of the topology of the free energy surface, five macrostates were defined with boundaries near transition regions in the $\phi-\psi$ parameters. States 1 and 2 span the regions of lowest free energy, usually designated C_{7eq}, and state 5 spans a high energy local minimum usually designated C_{ax}.

Using this state space definition, each conformation of each trajectory can be assigned a macrostate index. Consequently, each of the trajectories is represented as a time-ordered string of macrostates, e.g., ...,4,4,5,4,5,3,3,.... From these it is trivial to obtain the functions $\Omega^{(i)}(t)$, at multiples of the sampling period. This is represented as a sequence of ones and zeros indicating whether the state of the trajectory is in state i or not, where i is a macrostate index. For example, $\Omega^{(3)}(t)$ = ...,0,0,0,0,0,1,1,....; $\Omega^{(4)}(t)$ = ...,1,1,0,1,0,0,0,....; $\Omega^{(5)}(t)$ = ...,0,0,1,0,1,0,0,.... These functions are then used with eqs 4, 6, and 11 to produce Boltzmann weighted macrostate populations, correlation functions, and lifetime distributions.

Equilibrium populations for the five states are shown in Table 1. This table also shows the degree of consistency observed between the populations of conformations produced from the replica exchange simulations, those produced during the kinetic simulations, and those implied from various transition matrices.

Boltzmann weighted lifetime distributions for each macrostate were computed from the simulation data using eq 11. Of fundamental interest is the degree to which these distributions are consistent with those of a true Markov chain. We compared each of the observed distributions with ones that would be produced by a Markov chain having the same mean lifetime. Because lifetime distributions depend on the time interval between samples, this comparison was done at various temporal resolutions. This was done by using values of τ in eq 11 that were different multiples, n_{lag}, of the underlying sampling period of 0.5 ps, which has the effect of computing lifetimes based on data sampled at different intervals.

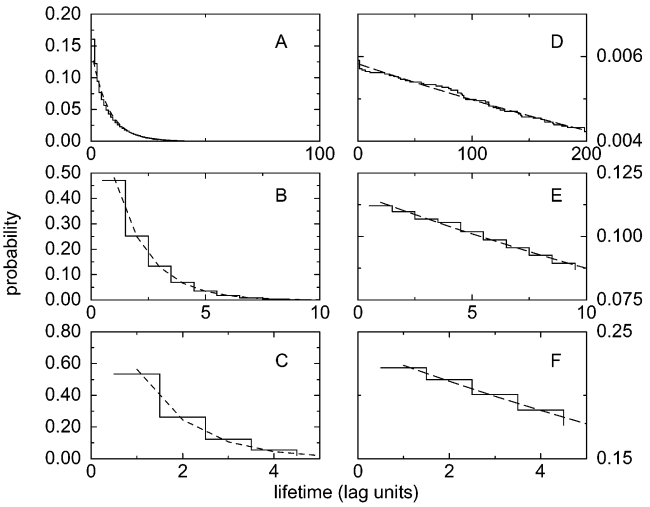

Figure 2. Lifetime distributions for macrostate 1 (left three panels) and macrostate 5 (right three panels), using three different lag times of 0.5 (top), 10 (middle), and 20 (bottom) ps, corresponding to lag times of 1, 20, and 40 sample periods. The lifetime distributions are in terms of lag time. The dashed lines on the left three panels show the distribution expected for a true Markov process with the same mean lifetime. The dashed lines on the right are exponential fits to the observed lifetime distribution for macrostate 5.

Representative lifetime distributions for macrostates 1 and 5 are shown in Figure 2. Along with the observed lifetime distributions for macrostate 1 the left side of the figure shows the distribution expected from a true Markov chain with the same mean lifetime. Macrostates 2−4 show behavior very similar to that of macrostate 1, suggesting that the behavior of these states is consistent with that of a Markov chain. The right side of the figure also shows the observed lifetime distribution for macrostate 5. Relative to the other states, macrostate 5 was very long-lived, with lifetimes that appear to be comparable to or greater than the 100 ps kinetics simulations. Because the distribution of lifetimes we observe for this state is limited to lifetimes of 100 ps or less, at which the distribution is still nonzero, it is inappropriate to compute a mean lifetime from these data for the purposes of comparison with a true Markov chain. Rather, for this state, we show an exponential fit to the observed data (all R values greater than 0.99).

The trajectory data were also used to compute Boltzmann weighted transition functions and matrices. For a system that is to be characterized with M macrostates, there will be M^2 correlation functions to be computed using eq 4, and, when these are normalized by use of the macrostate probabilities computed using eq 6, one obtains an equal number of transition functions. These functions were evaluated at discrete times, namely multiples of the sampling period of $\tau_{samp} = 0.5$ ps, from zero up to approximately $100\tau_{samp}$, or 50 ps. These functions have a simple characteristic look. Diagonal transition functions, $T_{ii}(t)$, represent the probability of finding the system in some macrostate i given that it was observed to be in the same state some time t earlier. These generally show a smooth decline from a value of unity. At infinite time, they would be expected to decay

TABLE 2: Transition Matrices for the Alanine Dipeptide at Three Different Lag Times

$$T(0.5\text{ ps}) = \begin{pmatrix} 0.8433 & 0.2196 & 0.3761 & 0.1279 & 0.0005 \\ 0.1559 & 0.7786 & 0.1094 & 0.7046 & 0.0002 \\ 0.0006 & 0.0002 & 0.4861 & 0.0203 & 0.0001 \\ 0.0002 & 0.0016 & 0.0261 & 0.1430 & 0.0003 \\ 0.0000 & 0.0000 & 0.0022 & 0.0042 & 0.9989 \end{pmatrix}$$

$$T(10\text{ ps}) = \begin{pmatrix} 0.5878 & 0.5781 & 0.5757 & 0.6076 & 0.0083 \\ 0.4102 & 0.4199 & 0.4159 & 0.3878 & 0.0086 \\ 0.0008 & 0.0009 & 0.0036 & 0.0000 & 0.0003 \\ 0.0010 & 0.0009 & 0.0024 & 0.0006 & 0.0003 \\ 0.0002 & 0.0002 & 0.0024 & 0.0040 & 0.9825 \end{pmatrix}$$

$$T(20\text{ ps}) = \begin{pmatrix} 0.5836 & 0.5843 & 0.5802 & 0.5595 & 0.0176 \\ 0.4142 & 0.4135 & 0.4106 & 0.4347 & 0.0163 \\ 0.0008 & 0.0009 & 0.0033 & 0.0006 & 0.0008 \\ 0.0010 & 0.0010 & 0.0020 & 0.0013 & 0.0003 \\ 0.0004 & 0.0004 & 0.0039 & 0.0039 & 0.9650 \end{pmatrix}$$

to the steady-state population for the state. The off-diagonal transition functions, $T_{ij}(t)$, represent the probability of finding the system in some macrostate i given that it was observed to be in some state j some time t earlier. These generally show a rise from zero to the steady-state population of state i.

From the M^2 transition functions computed at discrete lag times, one may construct transition matrices. There is an $M \times M$ transition matrix for each time at which the transition functions were computed. For the matrix associated with any choice of lag time, $t = n_{\text{lag}}\tau_{\text{samp}}$, each column sums to unity. The elements of column j describe the probability of observing the system to be in state i, given that it was in state j at time t earlier. Thus, the matrix can be thought of as characterizing the evolution of the system by a particular time increment, $t = n_{\text{lag}}\tau_{\text{samp}}$. Representative transition matrices are shown in Table 2 for lag times of 0.5, 10, and 20 ps.

The transition matrices that correspond to different degrees of observed temporal evolution have many of the properties of a Markov transition matrix. In particular they can be diagonalized and the eigenvalues can give information about the time scales over which the system changes state. Processes that strictly exhibit detailed balance will produce transition matrices with real eigenvalues. In our simulations, even with a time reversible dynamical integration algorithm, detailed balance is not strictly observed in the space of macrostates due to the finite duration of our simulations. Therefore, the eigenvalue spectrum produced from our transition matrices, though real and positive for most lag times, is occasionally complex. The real part of the eigenvalue spectrum is shown in Figure 3 as a function of the time index of the associated transition matrix. If the evolution of the system can be described by a Markov process, these curves exhibit simple exponential decay.

For a Markov chain characterized by some transition matrix, $T(t)$, that propagates the system by a time interval t, the time scale for exponential relaxation implied by any particular eigenvalue, μ, is given by $\tau_{\text{relax}} = -t/\ln \mu$. This function of the eigenvalues less than unity for each of the observed transition matrices is shown in Figure 4. If the system could be described as a Markov chain, this function would be constant. The feature to notice about these graphs is that the functions exhibit a rise at short times, corresponding to non-Markovian behavior, and then reach a plateau at longer times where they may be considered *approximately* Markovian. The upper panel in the figure shows the behavior of the time scale associated with the largest nonunity eigenvalue as a function of lag time. This

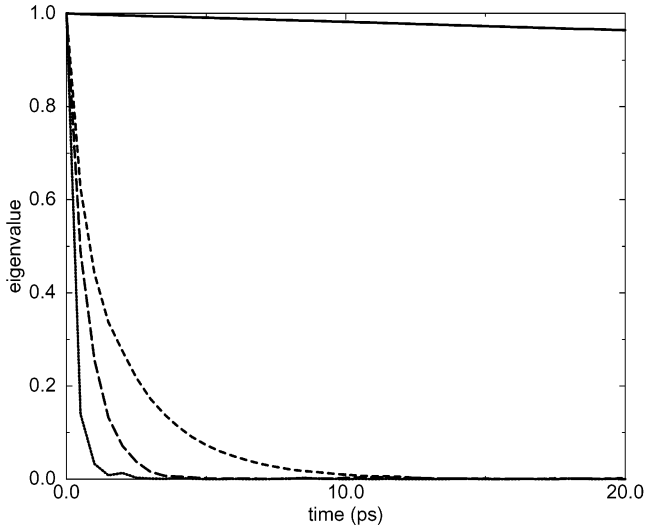

Figure 3. Four nonunity eigenvalues of the alanine dipeptide transition matrices that correspond to various amounts of temporal evolution.

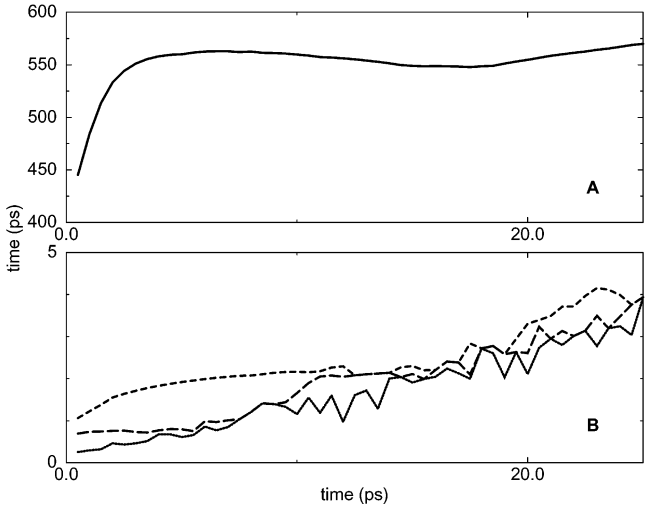

Figure 4. Time scales implied for the alanine dipeptide by eigenvalues of the observed transition matrices corresponding to various amounts of temporal evolution.

implies a time scale of approximately 550 ps for the slowest relaxation process in this system. Examination of the eigenvector associated with this slow process indicates that it involves transitions to and from state 5. The lower panel shows the behavior associated with the other three eigenvalues, which correspond to the fastest processes in the system. The largest of these shows a rise to a plateau of about 2 ps at lag times of about 10 ps. The two fastest processes have times of 0.5 ps or less. Processes with characteristic times that are comparable to or less than the 0.5 ps sampling period cannot be described. A interesting feature of these graphs is the slow and approximately linear rise after the plateau is reached. This occurs at lag times of about 10 ps for the fast time scales and after about 20 ps for the slow time scales. This feature is also observed when we analyze true Markov processes and is affected by the number and length of simulations used in the analysis.

The central result of this section is that the analysis of the alanine dipeptide system reveals roughly Markovian behavior on a 10−20 ps time scale, that the slowest relaxation processes occur on a time scale of approximately 550 ps, and that this information can be obtained from the analysis of many simulations that are significantly shorter than this time.

4. β-Hairpin System

4.1. Simulation Methods. The replica-exchange[39,40] thermodynamic simulations of the C-terminal β-hairpin from protein G in explicit water carried out by Zhou et al.[16] provided the starting point for this work. The equilibration protocol, protein system, and force field (OPLS-AA[45] with SPC[46] water) are all identical to that prior work, allowing us to use the thermodynamic data as a basis for our kinetic simulations. The replica-exchange data at 310 K included a total of 19 086 postequilibration conformations of the peptide. A sample of 287 of these conformations were selected as the starting conformations for microcanonical (NVE) kinetic simulations. Instead of simply selecting conformations from the replica-exchange data with uniform probability, which would have produced a correctly Boltzmann weighted sample of kinetics runs by construction, we deliberately biased the sampling toward regions of conformational space that we suspected were potential transition states or bottlenecks in the folding free energy landscape.

First, a range of order parameters were calculated for each peptide conformation in the replica-exchange data set. These included the properties calculated in the previous work, such as the radius of gyration (R_g), the number of native hydrogen bonds based on distance and angle criteria (HBcount), and the fraction of native contacts (ρ), as well as a range of new order parameters including the ϕ and ψ angles for each residue in the peptide, the distances corresponding to native and non-native salt bridges, and the van der Waals contacts between the core hydrophobic residues (Tyr6-Phe13, Trp4-Phe13, and Trp4-Val15). Selection criteria were then expressed in terms of these properties, defining *selection cells* in conformational space. All of the 310 K replica-exchange conformations that fell into a particular selection cell were recorded, and 10−20 of these conformations were randomly selected as the starting points for NVE runs. A total of 26 different selection cells were used, and approximately 10−20 conformations were drawn from each to yield a total of 287 starting states for kinetic trajectories. The criteria used for each selection cell differ, as well as their stringency—some "cells" include all of phase space by construction, whereas others only had a few (20−50) representatives in the replica-exchange data. The detailed criteria defining each selection cell are reported in Supporting Information. Initially, these criteria were used to bias sampling toward regions of the 2D energy landscapes (HBcount vs R_g, ρ vs R_g) where there appeared to be transition states or bottlenecks. Later, the selection cells were used to sample more heavily in regions of phase space that might be near the transition state for formation of the hairpin turn or the first critical native hydrogen bond. This deliberately biased sampling of the starting points for the NVE trajectories produces a non-Boltzmann sample that was subsequently corrected to a 310 K Boltzmann distribution through the selection cell formalism described in the companion paper. All reported averages, whether of kinetic or thermodynamic properties from these simulations, are the appropriately Boltzmann weighted averages for the ensemble of interest.

The IMPACT program package[47] was used to solvate and equilibrate each selected conformation prior to production microcanonical (NVE) simulation. The peptide was solvated by 1800 SPC[46] water molecules in a 38 Å box, along with three sodium counterions to make the entire solvated system electrically neutral. Five hundred steps of conjugate gradient minimization were performed using a finite ranged potential with a 9 Å cutoff. Throughout the equilibration, a molecule-based cutoff was used for the solvent and an atom-based cutoff for the solute. Also, all bonds were constrained to their equilibrium values using SHAKE[38] and RATTLE[48] with a tolerance of 10^{-7} Å. After minimization, the particle mesh Ewald (PME) technique[49] was used to treat all long-ranged electrostatic interactions. The solvent was equilibrated by six successive runs of canonical (NVT) molecular dynamics of 5 ps each where the target temperature was 60, 110, 160, 210, 260, and 310 K, respectively. Both Andersen[41] and Berendsen[50] thermostats were used. For the former, velocities were reassigned from the appropriate Maxwell−Boltzmann distribution every 100 time steps. A coupling constant of 0.01 ps^{-1} was used for the latter. During this solvent equilibration phase all of the protein atoms were constrained to their initial positions. The protein, ions, and nearest 1660 water molecules were retained for the next phase of the calculation and the excess water molecules were discarded, though the original box dimensions were retained. This yielded a total of 5239 atoms in each production system. This second system was reequilibrated using an identical molecular dynamics protocol to the one described above, except that in this phase the solute atoms were not constrained. The positions and velocities of the final conformation of this equilibration phase were directly used as the initial conditions of a production microcanonical (NVE) simulation.

Production microcanonical (NVE) simulations were run using the Blue Matter[32] program package, a parallel molecular dynamics application specifically designed for the IBM Blue Gene[33] research prototype hardware but which can also be used on standard commercially available computer hardware. Microcanonical NVE simulations used a velocity Verlet[51] integrator with a 1.0 fs time step. No temperature or pressure control was used, and the total energy drift averaged over all 287 trajectories was 0.0002 kcal mol^{-1} ps^{-1}, with a standard deviation of 0.0007. If this energy drift were to go entirely into kinetic energy, it would correspond to a temperature drift of less than 10^{-2} K.

All bonds to hydrogen were constrained using SHAKE[38] and RATTLE[48] with a tolerance of 10^{-8} Å. Electrostatic interactions were again treated with the particle mesh Ewald[49] algorithm. An atom-based switch function was used to truncate smoothly both the Lennard-Jones potential and the direct space term of the Ewald potential over a range from 9 to 10 Å. The PME reciprocal space calculation used a grid spacing of 0.5 Å with a fourth-order interpolation. The force arising from the reciprocal space potential was calculated analytically rather than by interpolation. Coordinates were sampled every 0.25 ps from each simulation. In general, each simulation was run for 0.5 ns, with some slight variation. This yielded a total of approximately 0.12 μs of kinetic simulation, producing over 500 000 conformations broadly distributed across the folding landscape of the hairpin peptide.

A number of different order parameters were calculated for each conformation from the 287 NVE simulations. For this work, we focused on one measure of overall collapse, the radius of gyration[52] of the core residues (R_g(core)), and on four different measures of hydrogen bond formation for the six native hydrogen bonds. The various hydrogen bond metrics differed in their functional form and in the strictness of their hydrogen bond definition. The most permissive metric was based on a simple measurement of the distance from the donor hydrogen to the acceptor atom. Histograms of these sorts of distances from the original replica-exchange data showed multimodal distributions with a major peak near 2.5 Å (corresponding to the hydrogen-bonded state) and a minimum between 5 and 6 Å. In some cases a second peak around 6.5−7.0 Å was observed, which appeared to correspond to the presence of a bridging water between the two groups (data not shown). Therefore, one

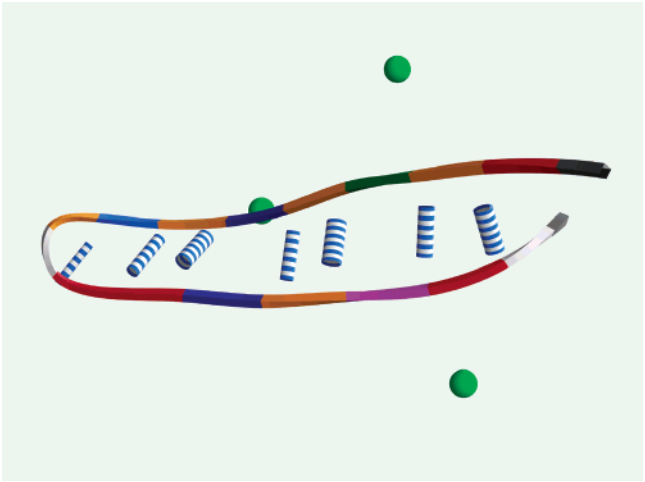

Figure 5. Representation of the nativelike conformation of the β-hairpin. Hydrogen bonds are indicated as cylinders with stripes, the diameter of the cylinder giving an indication of the strength of the hydrogen bond. Counterions are represented as spheres. In this work we are primarily concerned with the rightmost six native hydrogen bonds in the figure. In our notation, the hydrogen bonding pattern for this conformation would be indicated as 111111. The leftmost hydrogen bond in the figure is weak and transient because of the orientation of residues in the turn.

hydrogen bond metric, denoted DA5.5, was based on a distance of 5.5 Å for the donor hydrogen to the acceptor atom. A second geometrically based measure of hydrogen bond formation, denoted GEOM, used the common distance and angle based criteria. For this, a hydrogen bond was counted as formed if the donor−acceptor distance was less than 4.0 Å and the donor−hydrogen−acceptor angle was greater than 120°. The final two metrics both used the DSSP energetic criterion for hydrogen bond formation,[53] but with either a permissive (<-0.5 DSSP energy units) or a restrictive (<-1.5) threshold. These were denoted DSSP0.5 and DSSP1.5, respectively. For each of the four definitions, the presence or absence of the six native hydrogen bonds was determined for every peptide conformation.

4.2. Analysis and Results. States for the transition matrix analysis were chosen in a way that we hoped would allow extraction of information about the temporal order in which native hydrogen bonds were formed, as well as the overall time scale for the process of collapse and native structure formation. States were defined with respect to two order parameters: radius of gyration,[52] R_g(core), and the hydrogen bond status (i.e., either hydrogen bonded or not) of each of the six residue pairs that are hydrogen bonded in the fully folded conformation (Figure 5). This status can be indicated by an ordered string of six characters, each of which is either a 1 if the residue pair is hydrogen bonded, or a zero if the residue pair is not. The first character of the string represents the status of the residue pair closest to the termini of the strand; the last character represents the status of the pair closest to the turn of the native state. Thus, 000000 represents the set of conformations with no native hydrogen bonds formed, 111111 represents the set of conformations with all six formed, and 000001 represents the set of conformations with a single native hydrogen bond formed, the one closest to the turn. This characterization results in $2^6 = 64$ possible hydrogen bond states. In our data set, the most populated states were those with a hydrogen bond status of 000000, 000001, or 000010. These states were further subdivided on the basis of R_g(core) values. The conformations with hydrogen bond status 000000 were divided into four sets: $R_g(\text{core}) \leq 5.25$, $5.25 < R_g(\text{core}) \leq 7.5$, $7.5 < R_g(\text{core}) \leq 9.5$, and $R_g(\text{core}) > 9.5$, denoted S (small), M (medium), L (large), and E (extended), repectively. Those with hydrogen bond status 000001 and 000010 were each further divided into two sets: $R_g(\text{core}) \leq 5.25$ and $R_g(\text{core}) > 5.25$, denoted S (small) and M (medium), respectively. These boundaries are near minima in the R_g(core) probability distributions observed for each of the three hydrogen bond patterns for each of the four different definitions of hydrogen bond status. This resulted in 69 states. However, from observation of trajectories, it was noted that for states with any hydrogen bonding, the bond nearest the turn was transient, frequently forming and breaking with very short lifetimes. This was true even for conformations that were otherwise fully hydrogen bonded. Therefore, except for those with hydrogen bond patterns 000000, 000001, 000010 and 000011, all other hydrogen bond states were combined in pairs without regard to the status of the hydrogen bond near the turn. This means, for example, that a new state was formed from including all conformations that would be characterized as either 000110 or 000111 into a state denoted 00011X, X meaning that hydrogen bond can be either formed or not. This combining process reduced the number of states by 30, from 69 to 39.

Next, each conformation from each trajectory was classified with respect to these 39 states. This was done using each of the four definitions for the existence of a native hydrogen bond. After this classification it was noted that there were several states with extremely low populations, e.g., less than 200 observations out of over half a million conformations. The nature and number of these states depended on the choice of hydrogen bond definition. It was felt that for transitions involving these infrequently observed states, insufficient data existed for a very precise characterization of their associated transition probabilities, so these states were considered for lumping with other states. The decision to lump or not, and with which other state the lumping should be done, involved examination of trajectory information to see what other state(s) occurred immediately before and after the infrequently observed states. These are the states that are kinetically accessible in a short period of time (a single sampling period) to the infrequently observed one. In most cases there was a single such state, and it differed from the infrequent state by the addition or removal of one hydrogen bond. These two macrostates were merged, meaning that the macrostate spanning the less frequently observed region of phase space was deleted and the more frequently observed macrostate was redefined as spanning both regions of phase space. For example, in schemes DSSP0.5 and GEOM, the very rare states 11010X and 11100X were both kinetically accessible to the more prevalent state 11110X, so the these three states were merged into a single macrostate that would be indicated as 11110X. In some cases, however, there were multiple states that were kinetically accessible to a rare one, and the infrequently observed state appeared to be serving as a transitioning state between two or more states with larger populations. In these cases, the merge was not done. This process resulted in different sets of states for each of the four choices of hydrogen bond definition. Following this procedure and using the DA5.5 hydrogen bonding crition resulted in 22 states, using the GEOM criterion resulted in 25 states, and using the DSSP0.5 and DSSP1.5 criteria resulted in 25 and 35 states, respectively. This process produced the macrostates used for the remainder of the analysis. These states are described in the Supporting Information.

Using one of the hydrogen bond definition criteria, the resulting hydrogen bond pattern, R_g(core), and the lumping process, each conformation of each trajectory can be assigned

TABLE 3: Macrostates Developed for Each of the Hydrogen Bond Definitions Used (Denoted DSSP0.5, DSSP1.5, DA5.5, and GEOM)[a]

	DSSP0.5		DSSP1.5		DA5.5		GEOM	
	desc	Boltz(obs)	desc	Boltz(obs)	desc	Boltz(obs)	desc	Boltz(obs)
1	000000E	0.03(0.14)	000000E	0.03(0.14)	000000E	0.03(0.14)	000000E	0.03(0.14)
2	000000L	0.02(0.14)	000000L	0.02(0.14)	000000L	0.02(0.13)	000000L	0.02(0.14)
3	000000M	0.04(0.14)	000000M	0.05(0.18)	000000M	0.04(0.10)	000000M	0.04(0.15)
4	000000S	0.00(0.02)	000000S	0.01(0.04)	000000S	0.00(0.00)	000000S	0.00(0.02)
5	000001M	0.00(0.02)	000001M	0.00(0.01)	000001M	0.00(0.01)	000001M	0.00(0.02)
6	000001S	0.00(0.01)	000001S	0.00(0.01)	000011	0.01(0.07)	000001S	0.00(0.01)
7	000010M	0.00(0.02)	000010M	0.01(0.05)	000010M	0.00(0.01)	000010M	0.00(0.01)
8	000010S	0.00(0.01)	000010S	0.00(0.01)	00010X	0.00(0.00)	000010S	0.00(0.01)
9	000011	0.00(0.07)	000011	0.01(0.04)	00011X	0.04(0.08)	000011	0.01(0.07)
10	00010X	0.00(0.00)	00010X	0.03(0.02)	001000	0.00(0.00)	00010X	0.00(0.00)
11	00011X	0.11(0.09)	00011X	0.13(0.09)	00101X	0.00(0.00)	00011X	0.12(0.10)
12	00100X	0.00(0.00)	00100X	0.00(0.00)	00110X	0.00(0.00)	001000	0.00(0.00)
13	00101X	0.00(0.00)	00101X	0.00(0.00)	00111X	0.06(0.12)	00101X	0.00(0.00)
14	00110X	0.00(0.01)	00110X	0.01(0.02)	01000X	0.00(0.00)	00110X	0.00(0.01)
15	00111X	0.11(0.10)	00111X	0.10(0.06)	010111	0.01(0.00)	00111X	0.11(0.09)
16	0100XX	0.00(0.00)	01X00X	0.00(0.00)	011100	0.00(0.00)	010011	0.00(0.00)
17	01111X	0.16(0.10)	01001X	0.00(0.00)	01111X	0.15(0.11)	011X0X	0.00(0.01)
18	01011X	0.00(0.00)	01010X	0.00(0.00)	100000	0.00(0.00)	01011X	0.00(0.00)
19	011X	0.00(0.00)	01011X	0.01(0.00)	100111	0.02(0.00)	01101X	0.00(0.00)
20	01101X	0.00(0.00)	01101X	0.01(0.00)	101111	0.00(0.00)	01111X	0.17(0.10)
21	10001X	0.00(0.00)	01110X	0.01(0.01)	110111	0.03(0.01)	1000XX	0.00(0.00)
22	10011X	0.00(0.00)	01111X	0.18(0.08)	11111X	0.55(0.18)	111X1X	0.42(0.11)
23	11110X	0.00(0.00)	1000XX	0.00(0.00)			11110X	0.00(0.01)
24	10111X	0.00(0.00)	10X10X	0.00(0.00)			10111X	0.01(0.00)
25	11XX1X	0.45(0.12)	10011X	0.00(0.00)			11011X	0.00(0.00)
26			1X100X	0.00(0.00)				
27			10101X	0.00(0.00)				
28			10111X	0.02(0.01)				
29			11000X	0.00(0.00)				
30			11001X	0.00(0.00)				
31			11010X	0.00(0.00)				
32			11011X	0.03(0.01)				
33			11101X	0.02(0.00)				
34			11110X	0.01(0.01)				
35			11111X	0.27(0.07)				

[a] The columns headed "desc" give a shorthand description of the dominant types of conformations that make up each macrostate. The notation gives the hydrogen bond pattern, the characters "E" (extended), "L" (large), "M" (medium), and "S" (small) refer to different ranges of R_g(core). The columns headed "Boltz(obs)" give the Boltzmann probabilities for finding each macrostate and the fraction of the total number of observed conformations that were consistent with the macrostate definition.

a macrostate index. As previously described, the trajectories are then represented as time ordered lists of states. Then using eqs 4, 6, and 11, along with the selection cell reweighting scheme to produce Boltzmann weighted macrostate populations, one can compute correlation functions and lifetime distributions.

The first results of this process are shown in Table 3. There is a high degree of consistency among the four macrostate classification schemes in terms of the nature of the most populated macrostates. Also shown in this table are the Boltzmann probabilities for each macrostate and the fraction of the total number of observed configurations that were consistent with the macrostate definition. The fact that these two values are different is a reflection of the fact that the starting conformations were not Boltzmann weighted.

Boltzmann weighted lifetime distributions for each macrostate and for each macrostate definition scheme were computed from the simulation data using a version of eq 11 that takes account the bias introduced by the selection cell sampling procedure.[1] We compared each of the observed distributions with ones that would be produced by a Markov chain having the same mean lifetime. As described above, this comparison was done at various temporal resolutions by using values of τ in eq 11 that were different multiples, n_{lag}, of the underlying sampling period of 0.25 ps.

Representative lifetime results are shown in Figure 6, where the GEOM macrostate definiton scheme was used. This shows lifetime distributions for two macrostates: macrostate 1 (000000E), representing states with a large radius of gyration and no native hydrogen bonds; and macrostate number 15 (00111X), representing states with three or four native hydrogens bond formed near the turn of the hairpin. Macrostate 1 is a low probability state in this scheme, and has a relatively long mean lifetime. Macrostate 15 is a high probability state (11%) and has a relatively short mean lifetime. The figure shows the distributions for each of these two macrostates computed with three different temporal resolutions corresponding to $n_{lag} = 10$, 50, 200 of our fundamental sampling period. For comparison, also shown on each histogram in this figure is the lifetime distribution that would be expected from a true Markov chain having the same mean lifetime as the observed data. The distribution for macrostate 1 is seen to have a qualitatively different shape than that of a Markov chain. This is true at all three temporal resolutions. The distribution for macrostate 15, on the other hand, shows non-Markov characteristics at the shorter time scales but is in good agreement with the Markov distribution at the longer time scale. Although to different degrees, for both macrostates the observed lifetime distributions appear more like the Markov distributions at longer time scales. The differences are systematic. Compared with the Markovian distributions, the observed distributions always have enhanced probabilities for very short and very long lifetimes. In general,

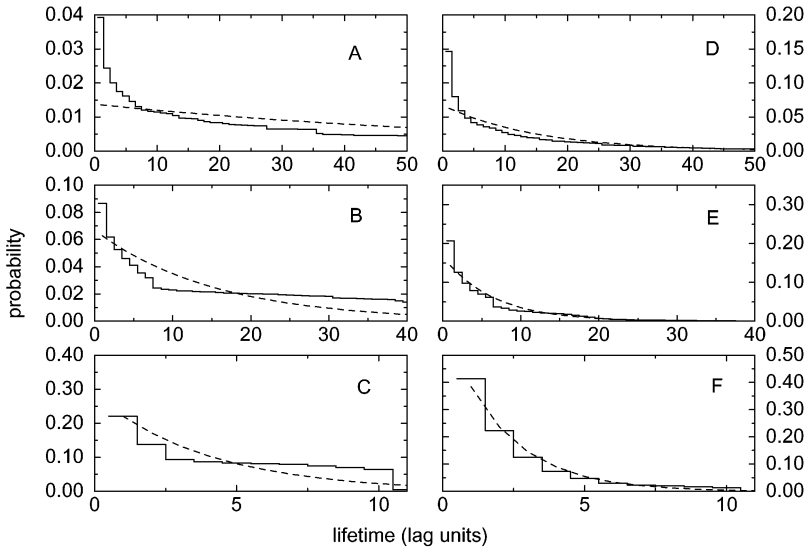

Figure 6. Lifetime distributions observed for two macrostates using the GEOM macrostate definition scheme. The three histograms on the left represent lifetime distributions for macrostate 1 (000000E) at three different temporal resolutions of (top, A) $10\tau_{samp}$, or 2.5 ps; (middle, B) $50\tau_{samp}$, or 12.5 ps; and (bottom, C) $200\tau_{samp}$, or 50 ps. The three histograms on the right (D, E, F) represent lifetime distributions for macrostate 15 (00111X) at the same three temporal resolutions. Shown on each histogram along with the observed (solid line) Boltzmann weighted distribution is the distribution with the same mean lifetime that one would expect if the process were truly Markovian (dashed line).

these trends are seen for all macrostates and for each of the four macrostate definition schemes.

The trajectory data were also used to compute Boltzmann weighted transition functions and matrices evaluated at discrete times, namely multiples of the sampling period of $\tau_{samp} = 0.25$ ps, from zero up to approximately $400\tau_{samp}$, or 100 ps. From the transition functions computed at discrete lag times, one may construct transition matrices. For the hairpin analyses, a particularly important feature of these transition matrices is that they are very nearly blocked into two submatrices, one of which indicates transitions among macrostates with no hydrogen bonding, and the other of which indicates transitons among macrostates with at least one hydrogen bond. The small matrix elements that connect these sets of states largely determine the slowest mode of relaxation, which corresponds to transitions from macrostates with no hydrogen bonds, to those with at least one.

The real part of the eigenvalue spectrum of the transition matrices, as a function of the time index of the associated transition matrix, is shown in Figure 7. If the evolution of the system could be described by a Markov process, these curves would exhibit simple exponential decay. The fact that this is not observed is another indication of non-Markovian behavior.

The time scale for exponential relaxation for a Markov process implied by any particular eigenvalue, μ, is given by $\tau_{relax} = -t/\ln \mu$. This function of the largest eigenvalue less than unity for each of the observed transition matrices is shown in Figure 8. If the system could be described as a Markov chain, this function would be constant, and clearly this is not the case. In the figure, two curves are shown for each macrostate definition scheme. One curve (solid) was produced using the trajectory data as generated. The other one (dashed) shows the effect of including both the forward and time-reversed version of each trajectory in the analysis. This procedure has the effect of *enforcing* detailed balance and results in all real eigenvalues for all of the observed transition matrices. One can see from the figure that this procedure affects the time scales implied by the observed transition matrices by factors of as much as three. Even more surprising, however, is the large difference in time scales implied by the use of different macrostate defintion schemes.

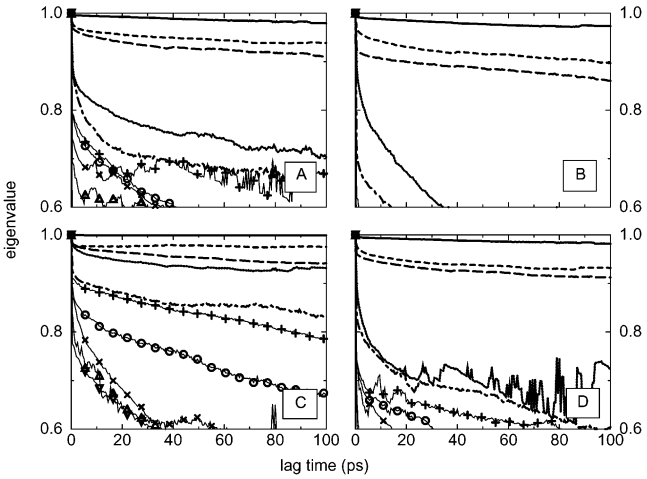

Figure 7. Largest eigenvalues of the transition matrices that correspond to various amounts of temporal evolution. The eigenvalues for macrostate definition scheme DSSP0.5 (A), DSSP1.5 (B), DA5.5 (C), and GEOM (D).

There are several indications that the observed transition matrices do not possess the properties one would expect of Markov transition matrices. Therefore, they should not be used as Markov transition matrices, especially to predict extremely long time behavior. However, in Figure 9 we show the result of such a process, anyway. Here, we have taken a particular transition matrix, constructed from transition functions evaluated at a time of $t = 200\tau_{samp} = 50$ ps, and repeatedly applied it to "evolve" a state started with probability only in macrostate 1. This is the macrostate with no native hydrogen bonds and a large radius of gyration in all four of the macrostate definition schemes. The initial state was propagated by 400 applications of the 50 ps matrix, implying an evolution of 20 ns. In each case we see an early and rapid reduction of the population of macrostate 1 (000000E), simultaneous with early and rapid growth in the populations first of macrostate 2 (000000L) then of macrostate 3 (000000M). This phase represents a rapid collapse to what might be called a molten globule state. In the DSSP0.5, DSSP1.5, and GEOM schemes, there appears to be an early buildup of probability in 00011X states before

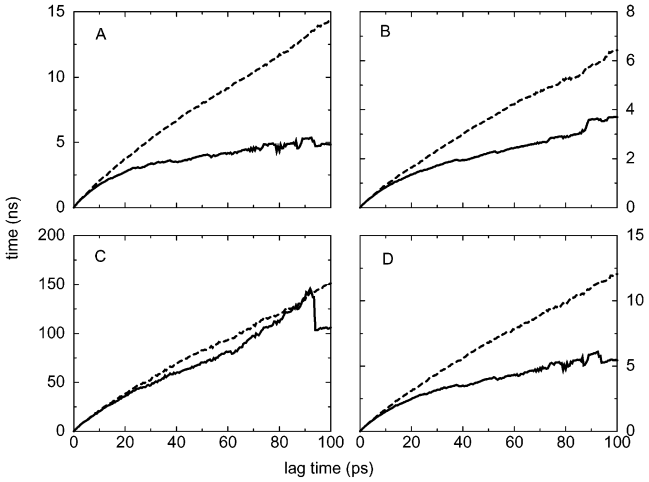

Figure 8. Times implied by largest eigenvalues of the observed transition matrices for the hairpin corresponding to various amounts of temporal evolution for macrostate definition schemes DSSP0.5 (A), DSSP1.5 (B), DA5.5 (C), and GEOM (D). For each macrostate definition scheme, there are two curves. The solid line represents the time scales predicted from using each trajectory once; the dashed line represents the same analysis using each trajectory *twice*, once forward, and once time-reversed.

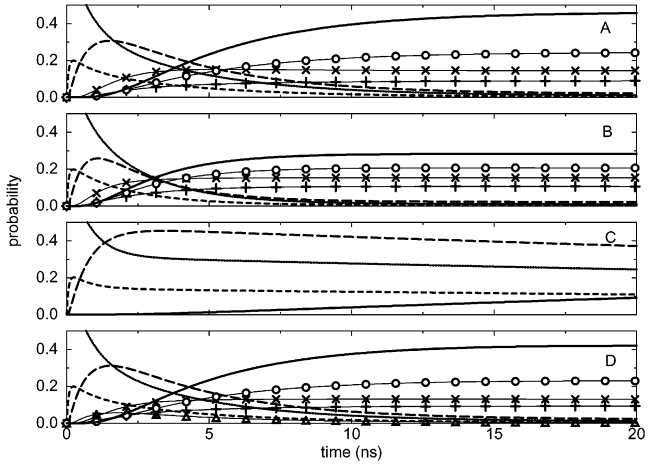

Figure 9. 20 ns evolution of the probability density implied by the use of the observed transition matrix corresponding to a lag time 50 ps. The initial probability density corresponded to all population in state 1 (000000E), the macrostate with no native hydrogen bonds and a large radius of gyration. This has been done for each of the four macrostate definition schemes: DSSP0.5 (A), DSSP1.5 (B), DA5.5 (C), and GEOM (D). In each panel, macrostate 1 (000000E), the monotonically decaying curve, is represented by the dotted line, state 2 (000000L) is represented by the line of short dashes, state 3 (000000M) is represented by the line of longer dashes, and the fully folded state is represented by the darker solid line. The states characterized as 00011X are represented by lines with X symbols [state 11 in schemes DSSP0.5 (A), DSSP1.5 (B) and GEOM (D)]. The states characterized as 00111X are represented by lines with cross symbols [state 15 in schemes DSSP0.5 (A), DSSP1.5 (B) and GEOM (D)]. The states characterized as 01111X are represented by lines with circle symbols [state 17 in scheme DSSP0.5 (A), state 22 in scheme DSSP1.5 (B) and state 20 in scheme GEOM (D)]. State 4 (000000S) is observed to play a role in the GEOM scheme and is represented by lines with triangle symbols.

probability grows for states with more native hydrogen bonds. This implies a *zipping* process, starting from the turn region. The behavior exhibited by the DA5.5 scheme is very different in that formation of the fully folded states does not seem to require a buildup of population in states with fewer numbers of native hydrogen bonds. Though not visible in the figure, there is a buildup of population in state 4 (000000S) that precedes the formation of the fully folded state in this scheme. The DA5.5 scheme uses a very permissive criterion for hydrogen bond formation where a native hydrogen bond is considered to be formed if the donor−hydrogen-acceptor pair are close enough that a water molecule probably cannot fit between them. In this scheme, it appears that native states arise from a fully collapsed state (000000S).

It is interesting that the folding process implied by the DA5.5 scheme seems to happen on a much longer time scale than that implied by the other schemes, because one would expect a more permissive hydrogen bond criteria to result in the appearance of faster and earlier folding. We feel that this may actually point to a problem with the other schemes. We believe that the more restrictive definitions of hydrogen bonding used in the other schemes results in the classification into the same macrostate of conformations that are very dissimilar. For example, a restrictive definition could place conformations where all the native hydrogen bonds are unformed but close to being formed into the same macrostate as conformations where all native hydrogen bonds are far from being formed. Then, evolution of the system between microstates that are structurally nearly native results in the appearance of transitions between the manifolds of states that are nearly native and those that have few or no native hydrogen bonds. The result is the production of transition matrix elements that imply artificially strong coupling between the macrostates and, thereby, unrealistically faster equilibration and relaxation times in the system. A more permissive definition of hydrogen bonding is less likely to lump into the same macrostate these kinds of kinetically distinct microstate. Inappropriate lumping of states is probably one cause of the non-Markovian behavior, as well as the fact that the folding time scales we predict from our simulations are 2−3 orders of magnitude faster than the 6 μs observed in experiments.[4,5]

Among the 287 trajectories sampled, many appeared to be stuck in the same macrostate for very long periods of time. In fact, if we restrict our attention to the longer trajectories of the set (the 227 that were at least 475 ps), there were 17 trajectories that were in a single macrostate at least 95% of the time according to *all four* of the macrostate definition schemes. The macrostates involved were always one of the three most extended states with no native hydrogen bonds. There were also several examples of trajectories that spent greater than 95% of the time in states with nearly a full complement of native hydrogen bonds, according to at least one of the macrostate definition schemes. In general, however, most trajectories explored a variety of macrostates. Because of their relatively short length, however, none was observed to evolve from fully extended states with no native hydrogen bonds to a fully folded state.

The trajectories where the system appears to be stuck in the same macrostate can be very useful. For our kind of analysis to work, the macrostates should be defined in a way that partitions phase space into regions that are kinetically homogeneous. Therefore, if we see situations where, in some trajectories, a macrostate appears to have a short lifetime, and, in others it appears to have extremely long lifetimes, there has apparently been an inappropriate lumping of conformations into the same macrostate. This indicates that the macrostate should be divided up into two or more smaller macrostates. Comparison of conformations from trajectories that are stuck with those from trajectories that are not can lead to the identification of new phenomenology that should be taken into account to produce a better macrostate definition scheme.

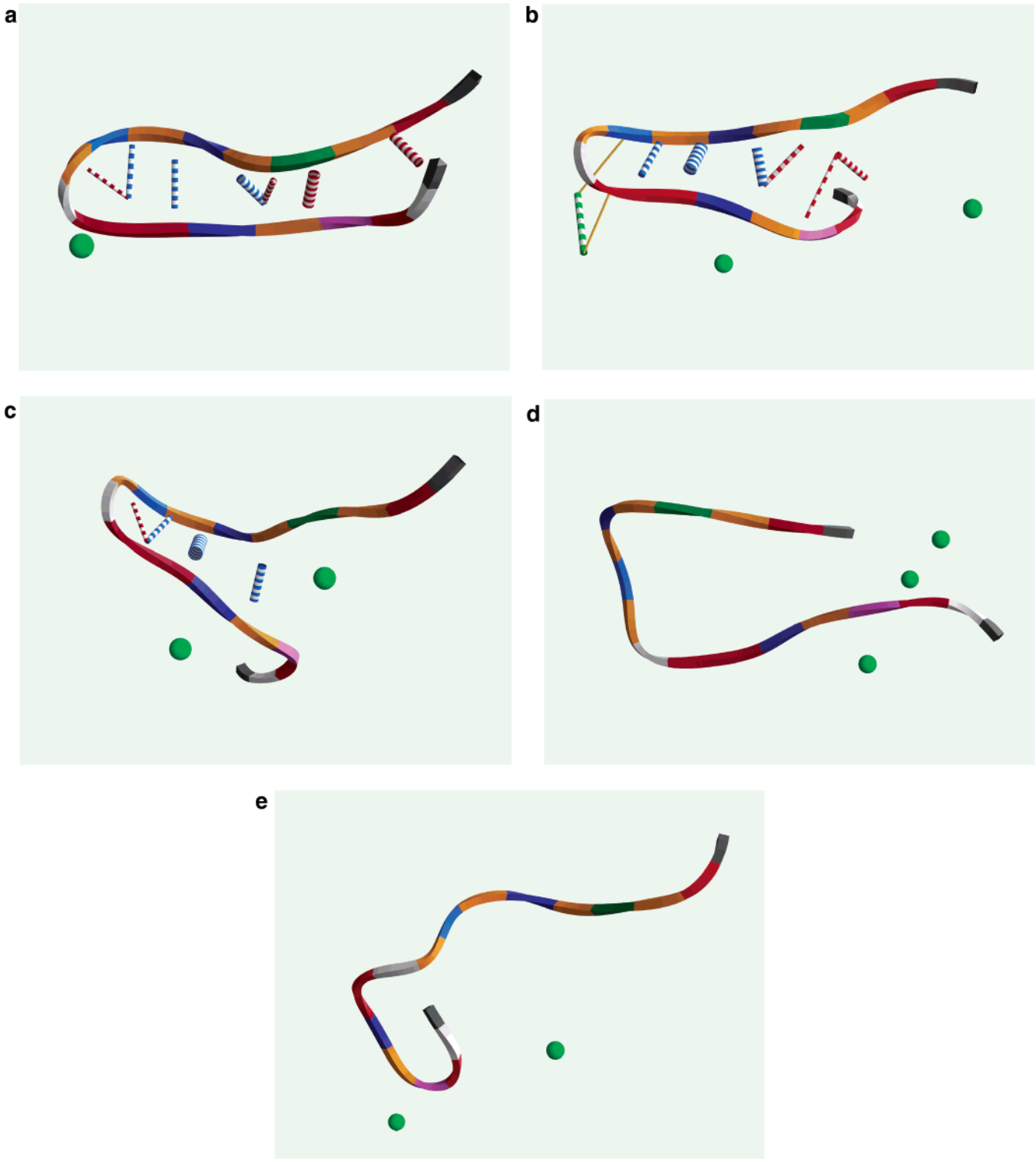

Figure 10. Selected conformations taken from trajectories in which the hairpin system stayed in the same macrostate for over 95% of the time. Panel A shows a conformation with four "misregistered" non-native hydrogen bonds that may be stabilizing this structure as well as a strongly associated ion. Panel B shows a conformation with three non-native hydrogen bonds, and with the terminal ends of the peptide splayed in a manner that might inhibit evolution. The cylinder in this panel near the turn of the hairpin indicates a properly formed salt bridge. Panel C shows a conformation with a non-native hydrogen bond near the turn. It is also splayed and has ions associated with residues near each of the termini. Panel D shows a conformation where the turn is misformed, making it impossible to form native hydrogen bonds. The open "loop" structure is stabilized by side-chain to backbone or inter-side chain contacts. Panel E shows a conformation where the C-terminal leg of the hairpin has folded back on itself in a tight turn formed by Trp and Glu residues.

Figure 10 shows some conformations taken from stuck trajectories. Panel A shows a conformation with three well-formed native hydrogen bonds; by most of the schemes this conformation would be considered to be in macrostate 000111. However, there are also four non-native hydrogen bonds that may be stabilizing this structure in a metastable conformation. The non-native bonds correspond to a misregistration of one strand of the hairpin relative to the other. There is also a strongly associated ion with this conformation that could be affecting the temporal behavior. Panel B shows a similar conformation

with three well-formed native hydrogen bonds, implying assignment to macrostate 000111, and three non-native hydrogen bonds. Additionally, the terminal ends of the peptide are splayed in a manner that might inhibit evolution of the conformation to other macrostates. The cylinder in this panel near the turn of the hairpin indicates a properly formed salt bridge. Panel C shows a conformation with three well-formed hydrogen bonds, with assignment to macrostate 000111, and a non-native hydrogen bond near the turn. This conformation is also splayed and has ions associated with residues near each of the termini. Panel D shows a conformation where the turn is misformed, making it impossible to form native hydrogen bonds. The open "loop" structure is stabilized by side-chain to backbone or inter-side chain contacts. Other features evident in this panel include Glu42 in contact with an ion (lower right), a persistent feature in the trajectory from which this configuration was taken. This conformation was assigned to macrostate 000000M by all four hydrogen bond definition schemes. Panel E shows a conformation where the C-terminal leg of the hairpin has folded back on itself in a tight turn formed by Trp and Glu residues. This turn structure persists for the entire 500 ps of the trajectory, but does not appear to be stabilized by any specific hydrogen bonding, hydrophobic, or ionic interactions. This conformation was assigned to macrostate 000000E by all four hydrogen bond definition schemes.

From these images one can infer the potential importance of non-native hydrogen bond formation, native and non-native side chain salt bridge formation, ion association, and conformations with splayed or twisted strands. We note that the work recently reported by Wei et al.,[21] also suggests the importance of non-native hydrogen bonds in the folding process. Order parameters used to construct macrostates to describe peptide folding rarely provide for these possibilities in the process. Therefore, when these kinds of conformations arise, they are lumped in with those that have very different temporal behavior. This seriously hinders any ability to understand the process in terms of Markov chains. However, it suggests that if these aspects of the process are addressed by inclusion of new criteria in the macrostate definitions, a Markov model of the process might be feasible.

5. Conclusions and Discussion

We have employed a rigorously derived set of formulas for the computation of transition probabilities from molecular dynamics data. The formulation uses Boltzmann weighted conformations as starting states for kinetic simulations and 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. 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 have applied this formalism to two example systems, an alanine dipeptide in a vacuum and the β-hairpin from Protein G in water. The alanine analysis used macrostates defined with respect to $\phi-\psi$ torsion angles, with boundaries obtained from examination of the free energy surface. It exhibits Markovian behavior on time scales longer than about 10−20 ps. The slowest relaxation processes in this system appear to be on the order of 550 ps, but the exact values for these times may be affected by the periodic velocity reassignments that were used to mimic the effect of a solvent. Regardless, these results show that our approach can be used to study the kinetics of conformational change in peptides, given sufficient sampling and adequate macrostate definitions.

The hairpin analysis used a novel macrostate space definition that resolves not only the number, but the pattern of native hydrogen bonds. We have tested four different criteria for hydrogen bonding. However, our analyses did not reveal Markovian behavior regardless of the hydrogen bond definition used. There could be many reasons for this. First, it is possible that the replica exchange simulations on which this study were based were insufficiently converged for us to deduce kinetics. The replica exchange simulations on the hairpin were relatively short and all replicas were started from a folded state. The resulting conformations could therefore be biased toward folded and slightly unfolded states. Our ensemble of starting states could therefore be missing conformations that are essential for characterizing the process of folding. This is a possible explanation for the absence of trajectories that cross from states with no native hydrogen bonds to states with at least one.

Second, our particular choices for the macrostate definitions may have involved inappropriate lumping of kinetically disparate states into the same macrostate. This has been discussed at length and it is clear how inappropriately lumped states can lead to the appearance of artificially fast kinetics as well as non-Markovian behavior. The examination of "stuck" trajectories provides valuable guidance in the formulation of better macrostates. It is clear from our examination of these trajectories that states may have to be defined that reflect not only the presence of native hydrogen bonds, but the presence of non-native hydrogen bonds, properly formed and improperly formed salt bridges, ion contacts, etc.

Third, in this demonstration of the method, we have performed rather short trajectories (approximately 0.5 ns). Recent experiments[54] on the dynamics of unfolded peptide chains have provided information on the end-to-end chain contact time as a function of chain length. These suggest that for peptides of the length of the β-hairpin it might be more appropriate to perform kinetic simulations of at least 5−10 ns.

We believe that better macrostate definition schemes and analyses done with more, better sampled starting states and longer simulations may show the emergence of Markovian behavior.

Because Markovian behavior is not strictly obeyed in our analysis, it is not appropriate to predict folding rates from the transition matrices we have computed. However, we can make a few qualitative statements about the β-hairpin folding process. Because few of our trajectories showed a significant degree of crossing between states with no native hydrogen bonds and those with at least one hydrogen bond, at the temperature of our study (310 K), there may be a large kinetic barrier between those two manifolds of states. This manifests itself in a transition matrix that is nearly blocked, with small off-diagonal elements connecting the blocks. These small off-diagonal elements determine the transfer of probability between these two manifolds and, thereby, the time scale for the formation of the native state.

It is notable that the time scales for folding implied by the various macrostate definition schemes (Figures 8 and 9) are much faster than what is observed experimentally. We feel this should not be of much concern, because our analysis clearly indicates that we have not observed behavior consistent with a Markov process, which is a prerequisite for predicting long time behavior. There are many possible explanations for this. In particular, slower processes would emerge with longer simulations, and it is not until we have observed the stability of observables with respect to simulation time that we would feel confident in predicting experimental folding rates. Longer and more simulations would be needed, for example, to better

characterize the transitions between the manifolds of states with and without hydrogen bonds. And, in agreement with the concerns of Fersht,[34] the process of characterizing these transitions may reveal other modes of behavior and other important folding pathways.

The mechanism of folding appears to depend on the criterion chosen for the existence of a hydrogen bond. Some hydrogen bond definitions imply that the pathway to folding from an unfolded state involves the formation of a native hydrogen bond near the turn of the hairpin, followed by a rapid zipping process. In this view, the time for folding is largely determined by the time it takes for the formation of the bond near the turn. On the other hand, a different hydrogen bond definition implies that the process involves a collapse, and that many hydrogen bonds then seem to form almost simultaneously, perhaps with the expulsion of water. Sensitivity of results with respect to hydrogen bond definition may explain some of the diversity of β-hairpin folding mechanisms proposed in the literature.

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 are looking for alternative assessment schemes to measure the degree to which Markovian behavior is observed, such as examination of the history dependence of the transition probabilities. We also wish to address the effect of using different force fields, and the sensitivity of our results with respect to the number and length of the dynamical simulations.[55] We also need to be careful that our analysis is based on an ensemble of uncorrelated and truly Boltzmann weighted starting states, from replica exchange simulations that have adequate sampling of phase space.

Properly applied, this approach has the potential to properly elucidate the kinetics 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.

Acknowledgment. We 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).

Supporting Information Available: List of macrostate groupings used in the analysis and tables of selection cells and order parameters. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem B* **2004**, *108*, 6571.
(2) Blanco, F. J.; Rivas, G.; Serrano, L. *Nat. Struct. Biol.* **1994**, *1*, 584.
(3) Blanco, F. J.; Serrano, L. *Eur. J. Biochem.* **1995**, *230*, 634.
(4) Munoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. *Nature* **1997** *390*, 196.
(5) Munoz, V.; Henry, E. R.; Hofrichter, J.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5872.
(6) Capaldi, A. P.; Radford, S. E. *Curr. Opin. Struct. Biol.* **1998**, *8*, 86.
(7) Eaton, W. A.; Munoz, V.; Thompson, P. A.; Henry, E. R.; Hofrichter, J. *Acc. Chem. Res.* **1998**, *31*, 745.
(8) Lapidus, L. J.; Eaton, W. A.; Hofrichter, J. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 7220.
(9) Pande, V. S.; Rokhsar, D. S. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9062.
(10) Zagrovic, B.; Sorin, E. J.; Pande, V. S. *J. Mol. Biol.* **2001**, *313*, 151.
(11) Dinner, A. R.; Lazaridis, T.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9068.
(12) Garcia, A. E.; Sanbonmatsu, K. Y. *Proteins* **2001**, *42*, 345.
(13) Kolinski, A.; Ilkowski, B.; Skolnick, J. *Biophys. J.* **1999**, *77*, 2942.
(14) Honda, S.; Kobayashi, N.; Munekata, E. *J. Mol. Biol.* **2000**, *295*, 269.
(15) Klimov, D. K.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2544.
(16) Zhou, R.; Berne, B. J.; Germain, R. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 14931.
(17) Zhou, R.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12777.
(18) Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 3555.
(19) Roccatano, D.; Amadei, A.; Di Nola, A.; Berendsen, H. J. C. *Protein Sci.* **1999**, *8*, 2130.
(20) Garcia, A. E.; Sanbonmatsu, K. Y. *Proteins* **2002**, *42*, 345.
(21) Wei, G.; Derreumaux, P.; Mousseau, N. *J. Chem. Phys.* **2003**, *119*, 6403.
(22) Snow, C. D.; Nguyen, H.; Pande, V. S.; Gruebele, M. *Nature* **2002**, *420*, 102.
(23) Bolhuis, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 12129.
(24) Phillips, J.; Zheng, G.; Kumar, S.; Kale, L. *Supercomputing 2002 Proceedings*, 2002; http://www.sc2002.org/paperpdfs/pap.pap277.pdf.
(25) Makino, J.; Taiji, M. *Scientific simulations with special-purpose computers*; John Wiley and Sons: Chichester, U.K., 1998.
(26) Shirts, M.; Pande, V. S. *Science* **2000**, *290*, 1903.
(27) Shirts, M. R.; Pande, V. S. *Phys. Rev. Lett.* **2001**, *86*, 4983.
(28) Folding@Home home page. http://folding.stanford.edu.
(29) Globus Alliance home page. http://www.globus.org.
(30) BlueGene project home page. http://www.research.ibm.com/bluegene/.
(31) Allen, F.; et al. *IBM Syst. J.* **2001**, *40*, 310.
(32) 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 Comput.* **2003**, *63*, 759.
(33) Adiga, N.; et al. *Supercomputing 2002 Proceedings*, 2002 http://www.sc-2002.org/paperpdfs/pap.pap207.pdf.
(34) Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 14122.
(35) 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.
(36) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E., III; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; et al. AMBER 6, University of California, San Francisco. http://www.amber.ucsf.edu
(37) Cornell, W. D.; Caldwell, J. W.; Kollman, P. A. *J. Chem. Phys.* **1997**, *94*, 1417.
(38) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.
(39) Hansmann, U. H. E., *Chem. Phys. Lett.* **1997**, em 281, 140.
(40) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60*, 96.
(41) Andersen, H. C. *J. Chem. Phys.* **1980**, *72*, 2384.
(42) Bolhuis, P. G.; Dellago, C.; Chandler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5877.
(43) Berne, B. J. *Chem. Phys. Lett.* **1984**, *107*, 131.
(44) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179.
(45) Jorgensen, W. L.; Maxwell, D.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.
(46) Walser, P.; Mark, A. E.; van Gunsteren, W. F. *Biophys. J.* **2000**, *78*, 2752.
(47) IMPACT, version 1.5, December 31, 2001; Molecular simulation software; Schrodinger, Inc.: New York, 1999.
(48) Andersen, H. C. *J. Comput. Phys.* **1983**, *52*, 24.
(49) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.
(50) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.
(51) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1982**, *76*, 637.
(52) We used a nonmass weighted radius of gyration based on the coordinates of the side chain heavy atoms of Trp43, Tyr45, Phe52, and Val54.
(53) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577.
(54) Krieger, F.; Fierz, B.; Bieri, O.; Drewello, M.; Kiefhaber, T. *J. Mol. Biol.* **2003**, *332*, 265.
(55) Chodera, J. D.; Dill, K. A. *Error Propagation And Sensitivity Analysis For Markov Models Constructed From Simulation Data*; UCSF Technical Report; University of California at San Francisco: San Francisco, CA, June 2003.