Folding probabilities: A novel approach to folding transitions and the two-dimensional Ising-model

Peter Lenz\textsuperscript{a)}
Lyman Laboratory of Physics, Harvard University, Cambridge, Massachusetts 02138
and Department of Chemistry, Stanford University, Stanford, California 94305

Bojan Zagrovic, Jessica Shapiro, and Vijay S. Pande
Department of Chemistry, Stanford University, Stanford, California 94305

(Received 3 November 2003; accepted 14 January 2004)

The theoretical concept of folding probability, \( p_{\text{fold}} \), has proven to be a useful means to characterize the kinetics of protein folding. Here, we illustrate the practical importance of \( p_{\text{fold}} \) and demonstrate how it can be determined theoretically. We derive a general analytical expression for \( p_{\text{fold}} \) and show how it can be estimated from simulations for systems where the transition rates between the relevant microstates are not known. By analyzing the Ising model we are able to determine the scaling behavior of the numerical error in the \( p_{\text{fold}} \) estimate as function of the number of analyzed Monte Carlo runs. We apply our method to a simple, newly developed protein folding model for the formation of alpha helices. It is demonstrated that our technique highly parallelizes the calculation of \( p_{\text{fold}} \) and that it is orders of magnitude more efficient than conventional approaches. © 2004 American Institute of Physics. [DOI: 10.1063/1.1667470]

I. INTRODUCTION

The problem of protein folding has become one of the main topics in statistical biophysics during the last years. While nowadays equilibrium properties of simple protein folding models are theoretically well understood,\textsuperscript{1} numerical simulations of the more detailed models required for direct comparison to experiments remain challenging.\textsuperscript{2} For example, the time scales reachable by conventional molecular dynamics (MD) simulations (nanoseconds to microseconds) are still orders of magnitude away from realistic folding times (tens of microseconds to seconds).

However, recently developed algorithms have lead to a novel approach to this problem, which seem to make the goal to numerically simulate the folding of realistic proteins achievable by using thousands of trajectories on the nanosecond time scale.\textsuperscript{3} The underlying idea is to investigate the system by not only one simulation but to start \( M \) statistically coupled simulations simultaneously.\textsuperscript{4} The coupling between the simulations is such that when one of these \( M \) simulations makes a transition to a new state, all other \( M - 1 \) simulations are stopped, reset to this new state and restarted. It can be shown that with this massive parallelization speedup factors close to \( M \) can be obtained. However, the main problem of this procedure is to decide which transient states should lead to such a reset of all replicas. Thus, a criterion is needed by which transitions between local minima of the free energy can be classified.

Such techniques allow for the simulation of protein kinetics and have quantitatively predicted protein folding rates.\textsuperscript{5} However, one also would like to use these simulations to understand how proteins fold. Towards this end, it is essential to identify the relevant degrees of freedom, i.e., the relevant “reaction coordinates” for folding.\textsuperscript{5}

That certain degrees of freedom are unsuitable as reaction coordinates is already evident in classic examples such as the gas to liquid transition of a van der Waals fluid. While the volume \( V \) clearly distinguishes gas from liquid, it turns out that \( V \) is a poor choice for the reaction coordinate. The kinetics of this system follows a nucleation mechanism, in which a nucleus of liquid phase grows in the gas. The transition state (i.e., the “critical nucleus”) for this system still has a gaslike volume. While \( V \) can distinguish gas from liquid, it fails to distinguish the gas from the transition state. Moreover, while there will be a free energy maximum in the volume, the configurations of the system associated with this maximum will not be the transition state. Thus, \( V \) has neither of the properties associated with a successful reaction coordinate. Here, a more appropriate choice is given by the size \( (R) \) of the nuclei. This more microscopic degree of freedom can distinguish the transition state from the initial and final states and the free energy maximum along \( R \) is associated with the true transition state.

This example is pertinent for several reasons: First, it demonstrates that not all degrees of freedom are suitable as reaction coordinates. Second, it illustrates that good reaction coordinates often make some connection to the kinetic mechanism. Finally, since nucleation is a dominant mechanism in biophysical chemistry (and indeed is typically considered a mechanism for complex phenomena such as protein folding), this particular instance may be a useful cautionary example of the pitfalls of choosing reaction coordinates.

Since there are many possible relevant degrees of freedom, one has to be able to decide computationally whether an arbitrary degree of freedom is suitable as a reaction coor-
eral analytical expression of the folding probability, \( p_{\text{fold}}(x) \), of a series of microstates (protein conformations) \( x \). By definition, if a state \( x \) has folding probability \( p_{\text{fold}}(x) \), then out of \( M \) simulations starting from configuration \( x \), \( M \) \( p_{\text{fold}}(x) \) end in the folded state. This quantity is able to capture the essential properties of the mechanism of folding kinetics: (i) \( p_{\text{fold}} \) can be used to identify the transition state conformations \( x_{\text{TS}} \), since they will be equally likely to fold or unfold [i.e., \( p_{\text{fold}}(x_{\text{TS}}) \approx 1/2 \); and (ii) degrees of freedom which correlate with \( p_{\text{fold}} \) should be suitable reaction coordinates.\(^7\)

While one can naturally calculate \( p_{\text{fold}}(x) \) computationally using many short simulations, this calculation can itself become quite burdensome, especially as the number of test conformations \( x \) increases. For this purpose we derive a general analytical expression of \( p_{\text{fold}} \) as function of (given) transition rates by solving a master equation. However, generally these transition rates are not known \textit{a priori}. We therefore demonstrate that they can be estimated from independent simulation runs. These analytical results can then be used to determine \( p_{\text{fold}} \) for a wide class of simulations. We give the following two examples: (i) for the two-dimensional Ising model; and (ii) for a simple protein folding model which we introduce here. In both cases we demonstrate that with our formalism the calculation of \( p_{\text{fold}} \) becomes considerably simpler and much less computationally demanding.

II. ANALYTICAL EXPRESSION FOR \( p_{\text{fold}} \)

To proceed, we consider a system with \( N_m \) microstates \( Y = \{y_i | 1 \leq i \leq N_m \} \). Every state \( y_i \in Y \) is characterized by its value \( \xi_i \) of the order parameter \( \xi \). We are interested in the transition between \( n_f \) folded and \( n_u \) unfolded states denoted by \( Y_f \subset Y \) and \( Y_u \subset Y \), respectively. For simplicity we order the states in such a way that \( Y'_u = \{y_i | 1 \leq i \leq n_u \} \) and \( Y'_f = \{y_i | N_m - n_f + 1 \leq i \leq N_m \} \). Note, that by definition all traps belong either to \( Y'_u \) or \( Y'_f \).

Generally, the transitions between these states will be described by a master equation,

\[
\dot{x}_i = \frac{dx_i(t)}{dt} = \sum_{k=1}^{N_m} K_{ij}x_j, \tag{1}
\]

where \( x_i \) is the density of state \( y_i \in Y \) in an ensemble of systems and \( (K_{ij}) \) is the \( N_m \times N_m \) rate matrix. Thus, \( K_{ij} \) is the rate (per unit time) of the transition from \( y_j \) to \( y_i \). One then has

\[
\sum_{i=1}^{N_m} x_i(t+\Delta t) = \sum_{i=1}^{N_m} x_i(t) + \sum_{i,j=1}^{N_m} K_{ij}x_j. \tag{2}
\]

\textit{Conservation of mass} implies \( K_{ii} = -\sum_{k \neq i} K_{ki} \). The folded and unfolded states are adsorption sites, i.e., \( K_{ii} = 0 \) for arbitrary \( j \) and \( y_j \in Y_f \cup Y_u \). Furthermore, \( K_{ii} = 0 \) for \( y_j \in Y'_f \cup Y'_u \) and \( K_{ii} < 0 \) for \( n_u < i \leq N_m - n_f \).

The folding probability \( p_{\text{fold}} \) of an arbitrary state \( y_m \) (with \( n_u < m < N_m - n_f \)) can be calculated by solving Eq. (1) with initial condition \( x_i = 1 \) for \( i = m \) and \( x_i = 0 \) otherwise. Then, \( x_j(\infty) = 0 \) for \( n_u < j < N_m - n_f \). By defining \( x^{(a)}(y_m) = \sum_{i=n_u}^{N_m} x_i(\infty) \) and \( x^{(f)}(y_m) = \sum_{i=n_u}^{m} x_i(\infty) \) one has

\[
p_{\text{fold}}(y_m) = \frac{x^{(f)}}{x^{(a)} + x^{(f)}}, \quad p_{\text{unfold}}(y_m) = \frac{x^{(a)}}{x^{(a)} + x^{(f)}}. \tag{3}
\]

To proceed, introduce a new matrix with elements \( L_{k+l} = K_{ij} \) where \( i = k + n_u \) and \( j = l + n_u \) with \( 1 \leq k,l \leq N_m - n_f - n_u \). Let \( \lambda_i \) denote the \( i \)th eigenvalue of \( L \), then \( R_{ij} = \lambda_i \delta_{ij} \), where \( \delta_{ij} \) is the Kronecker’s delta. The rows of the transformation matrix \( R \) are given by the eigenvectors of \( L \). Note, since \( L \) is not symmetric, the transformation matrix \( R \) is (in general) not unitary, i.e., \( R_{ij} \neq R_{ji}^T \). By integrating Eq. (1) one obtains \( x_i(t) = R_{i-j} \exp(\lambda_j t) R_{j-k} x_{k+n_u}(0) \) for \( n_u < i \leq N_m - n_f \) and \( t > 0 \). Thus,

\[
x^{(f)}(y_m) = \sum_{i=n_u}^{N_m} x_i(\infty) = \sum_{i=n_u}^{N_m} \int_0^\infty dt x_i+n_u(x)(t) = \sum_{i=n_u}^{N_m} \sum_{j=n_f-1}^{N_m} x_i+n_u R_{ij} \frac{1}{\lambda_j} R_{j-k} x_{k+n_u}(0). \tag{4}
\]

A similar expression can be derived for \( x^{(a)}(y_m) \).

Equation (4) is the main analytical result of this study. Below, we implement the following scheme: (i) we determine the rate matrix \( (K_{ij}) \) either by (a) analytically calculating the transition rates for a solvable model (Ising model), or by (b) analyzing numerical [Monte Carlo (MC)] data. (ii) We numerically determine the eigenvalues and eigenvectors of \( (L_{ij}) \) and numerically invert \( (R_{ij}) \). (iii) We calculate \( x^{(f)}(y_i) \) for all states \( y_i \) by using Eq. (4).

III. ISING MODEL

Before jumping to protein folding, a system with complicated kinetics, we first examine a simpler, well understood system to test our methodology: we consider a two-dimensional \( N \times N \) Ising model with Hamiltonian

\[
H = -J \sum_{(i,j)} s_i s_j. \tag{5}
\]

Generally, the phase space of the Ising model consists of \( N_m = 2(N^2) \) states. In order to be able to explore phase space in sufficient detail we restrict the numerical investigation to \( N=3 \). Although the number of microstates is then much smaller than that of any realistic protein folding model, the analysis of this model is quite useful since the exact values of the folding probability \( p_{\text{fold}} \) can be calculated. Thus, by comparing the exact with the estimated values it becomes possible to determine the scaling behavior of the error in estimating \( p_{\text{fold}} \) as function of the number of analyzed MC runs.

To implement the scheme described above we choose as order parameter the magnetization \( m_i \) of a spin configuration \( y_i \). The folded (unfolded) state is chosen to be the configuration where all spins are up (down). Thus, \( (n_u,n_f) = (1,1) \), \( m_1 = -N^2 \), and \( m_{N_m} = N^2 \). We order the states \( y_i \) by mapping the spin configuration \( y_i \) to a binary number \( b_i \) and
This system gives a good opportunity to demonstrate how generally the transition rates $K_{ij}$ can be estimated directly from MC data. For this purpose we performed $N_r$ MC runs, which all started from the unfolded configuration and stopped when the folded state was reached. From these $N_r$ MC trajectories, the transition rates could in principal be obtained via $K_{ij}^\text{est} = 1/\bar{\tau}_{ij}$ where $\bar{\tau}_{ij}$ is the averaged mean passage time (i.e., the number of MC steps) for the transition from $y_j$ to $y_i$. However, $\bar{\tau}_{ij}$ cannot be measured directly from the MC data. To illustrate this, consider a system consisting of four states $y_i$ with $i=1,2,3,4$ (in the following denoted by 1, 2, 3, and 4) and a MC trajectory given by, say, 11122111134. Naively, one would assign the following mean passage times:

$$
\bar{\tau}_{11} = 3,
\bar{\tau}_{12} = 2,
\bar{\tau}_{31} = 4,
\bar{\tau}_{43} = 1.
$$

However, this is wrong. In the first two steps of the sequence 1112 the system does not necessarily try to make a transition to 2 (it also could try to jump to, e.g., 3). The procedure given by Eq. (8) works only for the one-step transitions, i.e., for the transition 3→4.

In order to get a correct estimate for $\bar{\tau}_{ij}$ one first has to calculate the transition probabilities. Here, one finds

$$
\begin{align*}
 p_{21} &= 0.5, \\
p_{12} &= 1, \\
p_{31} &= 0.5, \\
p_{43} &= 1,
\end{align*}
$$

where $p_{ij}$ the transition probability of the transition $y_j \rightarrow y_i$. Then, the rates can be calculated via

$$
\frac{K_{ik}}{K_{jk}} = \frac{p_{ik}}{p_{jk}}.
$$

Thus,

$$
K_{ij} = \sum_j K_{ij}.
$$

Furthermore, $\sum_j K_{ij} = N_j/\bar{\tau}_{ij}$, where $N_j$ is the number of neighboring states of $y_j$, and $\bar{\tau}_{ij}$ is the mean waiting time of state $y_j$, i.e., the number of MC steps the simulations spends in state $y_j$ (which can be directly measured from the MC data). Thus, one finally finds

$$
K_{ij}^\text{est} = \frac{N_j p_{ij}}{\bar{\tau}_{ij}}.
$$

Another important point to note is that the MC simulations do not explore the full phase space, i.e., only $N_i < N_m$ spin configurations are reached. Correspondingly, in this case Eq. (12) does not yield an estimate for the full rate matrix but only for a $(N_i \times N_i)$ submatrix $K_{ij}^\text{est}$.

We have performed MC simulations to calculate 10000 trajectories (226 788 171 MC steps in total) and calculated

![Image](J. Chem. Phys., Vol. 120, No. 14, 8 April 2004 Folding transitions and the two-dimensional Ising-model 6771)

**FIG. 1.** The analytical folding probability $p_{\text{fold}}^\text{an}(y_i)$ of state $y_i$ as function of $p_{\text{fold}}^\text{mc}(y_i)$. Here, $p_{\text{fold}}^\text{mc}(y_i)$ is determined by averaging over $N_{\text{mc}}=1000$ MC runs with initial state $y_i$. Then sort the binary numbers $\{b_j\}$ according to their numerical values. Thus, $b_i < b_j$ for $i < j$ with $b_j = 1 + \sum_{i=1}^{N_j} 2^{j-1}(s_i + 1)/2$.

For this system the energy of every state is explicitly known and the transition rates for the MC dynamics can be calculated analytically. One has $K_{ij} = 0$ for all configurations $y_i$ and $y_j$ which differ by more than one spin flip and

$$
K_{ij} = \min\{1.0, \exp\left[\frac{(E_j - E_i)}{T}\right]\}
$$

otherwise. Here, $E_i$ and $E_j$ are the energies of $y_i$ and $y_j$, respectively. Additionally, we set $K_{ii} = 0$ and $K_{iN_{\text{mc}}} = 0$ for all $i$ to implement the special boundary conditions for the folded and unfolded state.

It is worth mentioning that here $|K_{ij}| \geq 1.0$ for all $y_i \in Y_F \cup Y_F$; since for all states $y_j$, a state $y_j$ exists which has lower energy. In particular, this implies that all eigenvalues $\lambda_i$ are real.

We analytically calculated $p_{\text{fold}}^\text{an}$ (within the scheme described above) and compared it with $p_{\text{fold}}^\text{mc}$ as obtained from MC runs. For this purpose $N_{\text{mc}}$ simulations were started from each spin configuration $y_i$. We choose $N_{\text{mc}} = 1000$ in order to keep the statistical error in determining $p_{\text{fold}}^\text{mc}$ as small as possible. Then, $p_{\text{fold}}^\text{mc}(y_i) = N_f(y_i)/N_{\text{mc}}$, where $N_f(y_i)$ is the number of MC runs starting from $y_i$ which find the folded configuration. As can be seen from Fig. 1 $p_{\text{fold}}^\text{an}$ and $p_{\text{fold}}^\text{mc}$ are (within the statistical error) identical. All shown results are for $N = 3$ and $J/T = 0.7$ (which is below the transition temperature).

In order to test the significance of the correlation in a more quantitative way, one can calculate the correlation coefficient $r(p_{\text{fold}}^\text{an},p_{\text{fold}}^\text{mc})$ for the data shown in Fig. 1. Here,

$$
r(x,w) = \frac{\sum_{i=1}^{N_m} (x_i - \bar{x})(w_i - \bar{w})}{\sqrt{\sum_{i=1}^{N_m} (x_i - \bar{x})^2 \sum_{j=1}^{N_m} (w_j - \bar{w})^2}},
$$

where $x = (x_1, \ldots, x_{N_m})$, $w = (w_1, \ldots, w_{N_m})$, $\bar{x} = 1/N_m \sum_{i=1}^{N_m} x_i$, and $\bar{w} = 1/N_m \sum_{i=1}^{N_m} w_i$. One then finds $r(p_{\text{fold}}^\text{an},p_{\text{fold}}^\text{mc}) = 0.998$ indicating that the two quantities are very strongly correlated.
the folding probabilities \( p_{\text{fold}}^\text{est} \) by analyzing subsets consisting of \( N_r \) trajectories. We choose \( N_r = 50, 100, 500, 1000, 5000, 10,000 \). Then, \( N_r = 297, 325, 414, 447, 500, 511 \), respectively. In Fig. 2 the folding probabilities \( p_{\text{fold}}^\text{est} \) obtained from \( K_{\text{est}} \) is shown for two sets of MC runs with \( N_r = 100 \) and \( N_r = 500 \), respectively. Then, \( K_{\text{est}} \) is a \( (325 \times 325) \) matrix and a \( (414 \times 414) \) matrix, respectively. This data is compared with \( p_{\text{fold}}^\text{an} \) calculated from the reduced \( (N_i \times N_i) \) rate matrix \( \tilde{K}_{ij} \) obtained from \( K_{ij} \) [given by Eq. (6)] by deleting those entries corresponding to states which are not reached by the MC trajectories. As can be seen from direct inspection of the figure, the agreement is not too good in particular for large under-sampling (i.e., small number of MC runs). There are several possible causes for these deviations: (i) \( p_{\text{fold}}^\text{an} \) is compared with \( p_{\text{fold}}^\text{mc} \) a quantity which itself might deviate from the exact value of \( p_{\text{fold}}^\text{est} \); (ii) the statistical error is large (in particular for small \( N_r \)); and, as mentioned, (iii) not all possible spin configurations are reached by the MC runs.

While we only can speculate about the relevance of (ii) and (iii), hypothesis (i) can be tested by comparing \( p_{\text{fold}}^\text{an} \) and \( p_{\text{fold}}^\text{est} \) with the exact values of the folding probability \( p_{\text{fold}}^\text{exact} \), see Fig. 3. Here, \( p_{\text{fold}}^\text{exact} \) is calculated (as in Fig. 1) by using the full \( (N_m \times N_m) \) rate matrix \( K_{ij} \) given by Eq. (6). As a comparison between Figs. 2 and 3 shows, in the latter case the agreement between the two quantities is better (although the correlation coefficients for both datasets are nearly identical). However, deviations still occur which might be due to the points (ii) and (iii) mentioned above. This is supported by the fact that results improve with larger \( N_r \), cf. Fig. 3.

These results already show that good estimates for \( p_{\text{fold}}^\text{est} \) can only be obtained by analyzing many MC simulations. However, since the exact value of the folding probability is known for the Ising model we can use these results to calculate the error of \( p_{\text{fold}}^\text{est} \) as function of the number of MC runs \( N_r \). In Fig. 4 the standard deviation \( \sigma \) of \( p_{\text{fold}}^\text{est} \) from \( p_{\text{fold}}^\text{an} \) is plotted, where

\[
\sigma^2 = \frac{1}{N_s} \sum_{i=1}^{N_s} (p_{\text{fold}}^\text{est}(y_i) - p_{\text{fold}}^\text{exact}(y_i))^2. \tag{13}
\]

By fitting the measured data we find that the standard deviation of \( p_{\text{fold}}^\text{est} \) scales as \( 1/(3N^{0.21}) \). As shown in Fig. 4 the error is, in the relevant range of \( N_r \), only slightly larger than that of the method of Du et al., where \( N_r \) simulations are started in configuration \( y \) to calculate \( p_{\text{fold}}(y) \) (leading to \( \sigma \sim N^{-1/2} \)). This demonstrates clearly the big advantage of our new approach: to obtain an accuracy of, say, 10% a direct calculation of \( p_{\text{fold}}(y) \) requires 100 simulations for every state \( y \), while we are able to achieve this goal for 400 states simultaneously by analyzing only 500 MC runs.
IV. FOLDING OF AN ALPHA HELIX

Next, we demonstrate how our method can be applied to protein folding. For this purpose we study a simple MC model for formation of alpha helices which combines an atomistic description of the peptide in implicit solvent with detailed energetics and a simplified move set. Even though alpha helices are some of the simplest structural elements in protein molecules, simulating their folding process in atomistic detail on an ensemble level is still extremely challenging from the computational perspective. However, in order to test the method of folding probabilities introduced above, we needed a large number of complete folding trajectories. With the help of our model we were able to accomplish this using a simplified MC scheme: we obtained hundreds of complete folding trajectories by using only conventional computational resources. The fact that our results match in many ways the results obtained by molecular dynamics and other types of simulation supports the validity of our simple move set and it motivates the use of our model in testing the concept of folding probabilities as introduced above.

For the sake of clarity and completeness we explain now our MC model in some detail. Below, we first describe the simulation method, and then compare our results with some of the more exact, but computationally more expensive methods. This will not only set the application of our method to calculate $p_{\text{fold}}$ on a solid basis but also demonstrates that the application of a very simple MC move set combined with an all atom representation and detailed energetics can yield results that are comparable to the ones obtained by more detailed simulation approaches.
A. Simulation method

Peptide: We have simulated the folding trajectories of a 21 residue polyalanine peptide that was capped at termini (NH$_2$ at the C terminus, and CH$_3$COO at the N terminus). The peptide was represented in atomistic detail with the unified atom formalism for methyl groups.

Move set. A single MC step consisted of a backbone move of a randomly chosen residue. Each move was accepted in accordance with the Metropolis criterion. The backbone moves were of two kinds. With probability of 10%, a backbone move consisted of randomly changing the $\Phi$, $\Psi$, and $\omega$ angles of a residue to values characteristic of either an alpha helix ($\Phi = -62^\circ, \Psi = -41^\circ, \omega = 180^\circ$), a 3–10 helix ($\Phi = -57^\circ, \Psi = -70^\circ, \omega = 180^\circ$), or a pi helix ($\Phi = -119^\circ, \Psi = 113^\circ, \omega = 180^\circ$) or an antiparallel beta sheet ($\Phi = -139^\circ, \Psi = 135^\circ, \omega = -180^\circ$). Alternatively, with probability of 90%, a backbone move consisted of changing the value of either the $\Phi$ or the $\Psi$ angle to a value drawn from a Gaussian distribution centered at the value of the angle under consideration with a standard deviation of $45^\circ$.

Overall, the frequency of attempted backbone moves for the $\Phi$ angles was identical to the frequency of attempted $\Psi$ angle moves. We have performed a series of 10 ns Langevin dynamics runs on a capped Ala-Ala-Ala tripeptide in order to calibrate the relative frequencies of different types of torsion angle updates. The Langevin dynamics simulations were run under the same conditions as the MC simulations (see below). The viscosity of the solvent was $\gamma = 91$ s$^{-1}$, the integration step was 2 fs, and no cutoffs were used. The RATTLE algorithm was used to constrain bonds. Based on these true-dynamics runs, we have calculated time correlation functions for different torsion angles of interest, and have furthermore calculated the values of the corresponding torsional diffusion coefficients based on the equation

$$
\langle v(t)v(0) \rangle = 3kT \exp\left(-\frac{kT}{mD}t\right),
$$

where $v(t) = d\theta(t)/dt$ is the torsional velocity at time $t$, $k$ is the Boltzmann constant, $T$ is the temperature, $m$ is the mass of a particle, $D$ is the effective torsional diffusion coefficient of interest, and $\langle f \rangle$ denotes a thermal average of $f$. We find that the $\Phi$ and $\Psi$ angle diffusion coefficients are quite similar to each other (within a factor of 2.5). Based on this, we expect that the same Monte Carlo move set and the same update frequency for the $\Phi$ and $\Psi$ angles, as introduced above, would be approximately consistent with their real time dynamics.

Folding simulations: All simulations were started from an extended conformation with all backbone and sidechain torsion angles set to $0^\circ$. The temperature was 300 K in all simulations. The OPLS force field with implicit GB/SA solvent was used with the Tinker Molecular Modeling package (http://dasher.wustl.edu/tinker) for energy calculations. Each simulation was run until at least 15 residues (not necessarily consecutive ones) adopted alpha helical $\Phi$ and $\Psi$ backbone values ($-72^\circ < \Phi < -54^\circ$ and $-54^\circ < \Psi < -36^\circ$) or until 500,000 MC steps have elapsed. We were able to simulate approximately 1,000,000 MC steps in a day on a single 550 MHz Pentium III processor.

B. Results and validation of the simulation method

The distribution of folding times for the alanine peptide (Fig. 5) is characterized by approximately single exponential behavior, in agreement with previous experimental and theoretical studies. The mean number of MC steps to fold the polyalanine peptide was $7.5 \times 10^4$. Note that the cumulative distribution function of the folding times exhibits somewhat sigmoidal behavior, which usually indicates the presence of an extra step preceding the transition of interest. In our case, the sigmoidicity may simply be a consequence of the nature of our starting configuration. Since our simulations start from a configuration with all of the torsion angles set to $0^\circ$, the peptide first needs to relax to a more favorable unfolded configuration before it folds.

Analysis of nucleation events (Fig. 6), defined as the appearance of three consecutive helical residues flanked by nonhelical residues, revealed several interesting trends. First, there is a significant tendency for the peptide to nucleate at the termini: the total number of nucleation events at residues 2 and 20 is more than three times larger than the mean number of nucleation events at residues 15.
Folding transitions and the two-dimensional Ising-model

Once nucleation occurs, there is a significant discrepancy between the propensity for helix propagation at the two ends of a helical stretch of residues. We have analyzed occurrences of helical additions to pre-existing helical stretches (defined as three or more residues with alpha helical backbone angles, see Sec. IV A): an extension was approximately 30% more likely to occur at the N terminal side of a helical stretch as opposed to its C terminal side (10364 N-terminal additions versus 8015 C-terminal additions). Considering that these numbers do not account for the inherent difference in nucleation propensities for the N and C termini of the peptides (we know that nucleation is more prevalent at the N terminus), it is likely that the discrepancy between the N and C extensions is even greater than 30%. A qualitative examination of the representative folding trajectories in Fig. 7 shows the apparent prevalence of N-end extensions. However, a physical and mechanistic model which would explain this result has yet to be determined. A potential clue may lie in our result showing that the residues in the N and C terminal regions undergo by far the largest number of transitions overall, which would clearly increase their odds of becoming helical relative to other sites. This result is in agreement with the more detailed MD simulations of polyalanine helix unfolding17,18 which showed that the variance of the values of backbone angles is largest at the termini of the helix, suggesting fraying at the ends. Major changes of the backbone angles in the middle of the peptide are more likely to result in significant steric clashes and other unfavorable interactions than are similar changes close to the termini of the peptide; these types of moves are therefore more likely to be rejected at sites in the middle of the peptide than at sites near the termini. The observed directional preference in helix propagation is in accordance with the MD results of Young and Brooks, who found that helix propagation is less favorable at the C terminus of a pre-existing helical stretch than at its N terminus.19 Their analysis of folding energetics indicated that this is due to the dominance at the C terminus of a helical stretch of the peptide–solvent interactions that favor the unfolded state over the folded state, which is stabilized by peptide–peptide interactions and peptide–solvent entropic contributions. Note, however, that this analysis dealt with propagation only and did not extend to nucleation.

Several other qualitative aspects of the folding process are illustrated in Fig. 7. For instance, in approximately half of all runs we observe more than one folding nucleus at distinct positions in the sequence; sometimes multiple nuclei propagate and merge into longer helical stretches. The effect of nucleation on the subsequent propagation is also important to note. Most theoretical models of the helix to coil transition (beginning with the earliest Zimm–Bragg model20) postulate that nucleation is the rate limiting step for folding, and that once a nucleus is formed, the helix propagates quickly. In agreement with this, we observe that nucleation in general does enhance the rate at which new helical residues are added. However, in an appreciable number of runs, the number of MC steps before a nucleation event is comparable to or even lower than the number of steps between additions of new helical residues to pre-existing helical stretches. Finally, we often observe isolated helical residues which are not flanked by helical residues on either side, but which remain helical for significant amounts of time. The curious stability of these residues can probably be assigned to stabilizing interactions, most likely nonhelical hydrogen bonds, between these residues and other residues further away in the sequence.

The energetics of the folding process is described in Fig. 8, which shows the total energy as a function of the number of MC steps. In most runs the total energy initially decreases rapidly and then tends to stabilize and remain roughly constant for the rest of the run.

Another important issue that needs to be discussed here is the applicability of Monte Carlo simulations for studying dynamic and kinetic aspects of protein folding. MC simula-
tions have typically been used to study thermodynamics and, in general, equilibrium behavior of complex systems. The questions we are addressing in this study, however, deal with time-dependent aspects of helix formation. For simple lattice simulations of folding it has been shown that the Monte Carlo algorithm gives an accurate description of folding kinetics on the ensemble level. In addition, Rey and Skolnick have demonstrated that a MC simulation and a Brownian dynamics simulation yield very similar folding pathways for an alpha helical hairpin. Furthermore, Shakhnovich and co-workers have recently analyzed the kinetic aspects of folding of the crambin molecule using Monte Carlo moves and a simple Go-like potential set. Their justification for using Monte Carlo to study folding kinetics is also applicable in our case. Briefly, if kinetic events of importance are separated by a sufficient number of local MC moves and are analyzed on an ensemble level, then they will reflect the properties of the underlying free energy landscape, and therefore, capture the essence of the real time kinetics. The kinetic description obtained in such a way is by necessity coarse grained, but is nevertheless realistic. In addition, we have approximately calibrated our Monte Carlo move set using more realistic Langevin dynamics simulations. The central idea of our calibration scheme was to adjust the relative frequencies of attempted Monte Carlo moves for different torsion angles in proportion with the torsional diffusion coefficients for these angles obtained in the Langevin simulations. Albeit coarse, this calibration scheme further assures that the ensemble averaged picture of the dynamics that we obtain is fairly realistic.

The move set in our simulations is partially based on the backbone $\Phi$ and $\Psi$ angles characteristic of five main types of secondary structure. The validity of using backbone $\Phi$ and $\Psi$ angles (instead, for instance, of hydrogen bonding patterns) as indicators of secondary structure is an important issue and has been well discussed before. The crux of the argument centers around the fact that the definition of the helix based on backbone angles yields results that are in better agreement with circular dichroism (CD) and nuclear magnetic resonance (NMR) experiments, the two primary means of studying helices experimentally. Finally, in our case, the use of backbone angles as structure determinants was a natural choice since it allowed us to create a simple move set. It should be noted that our move set is similar to the MC move set introduced by Srinivasan and Rose in their studies of secondary structure propensities of different sequences. It is a “smart” move in the sense that it steers the peptide towards adopting one out of several secondary structure conformations. This approach can be justified when dealing with sequences that one knows should adopt a particular secondary structure. The correspondence between our results and the results of other, more detailed simulations speaks in favor of this.

Finally, it is important to note that because of the lack of experimental data at this level of resolution, some of the details of our simulation cannot be directly verified as of now.

C. Determination of $p_{\text{fold}}$

Having shown that this model is able to capture some essential features of the formation of alpha helices, we now use the simulation trajectories to calculate $p_{\text{fold}}$ for the protein confirmations. In order to parametrize phase space details of the spatial configuration of the peptides can be ignored and only the position of alpha helices has to be stored. Consequently, configurations $y_i$ of the protein can be stored as binary numbers: $y_i = (r^{(1)}_i, \ldots, r^{(21)}_i)$, where $r^{(j)}_i = 1$ if residue $j$ belongs to an alpha helix and $r^{(j)}_i = 0$ otherwise. As order parameter serves $\xi(y) = \sum_{i=1}^{N_{\text{res}}} p^{(i)}_j$, i.e., the number of residues belonging to alpha helices. Finally, we set $Y_{\text{rel}} = \{y_i \in Y | \xi(y_i) = 15\}$ and $Y_F = \{y_i \in Y | \xi(y_i) = 3\}$ (as described above). Since the energy $E_j$ of a configuration $y_j$ is not explicitly known, the rate matrix $K$ has to be determined directly from the MC data.

Out of a total of 526 MC runs, $N_r = 466$ found the folded state. For our analysis only the completed runs were taken into account. In these runs $N_r = 66411$ out of $2^{21}$ possible configurations were reached by MC trajectories.

The calculation of $p_{\text{fold}}$ involves the inversion of the rate matrix. Due to computational limitations the number of degrees of freedom had to be reduced. We therefore split the $N_r$ states into a subset of relevant $Y_r \subset Y$ and irrelevant $Y_{\text{irr}} \subset Y$ configurations by requiring that for $y_i \in Y_r, (y_i \in Y_{\text{irr}})$ configuration $y_i$ occurred more than (less than) $N_{\text{rel}}$ times in the MC data. We chose $N_{\text{rel}} = 2000$ and ended up with $N_{\text{rel}} = 3239$ relevant states.

Next, we determined $K_{\text{est}}$ by analyzing the MC data as described above. In order to do so, irrelevant states had to be projected onto relevant states. Thus, a transition $y_i \rightarrow y_j \rightarrow y_k$ is interpreted as $y_i \rightarrow y_k$ for $y_i, y_k \in Y_r$, and $y_j \in Y_{\text{irr}}$. In Fig. 9 we present the folding probabilities $p_{\text{fold}}^{\text{est}}$ obtained from $K_{\text{est}}$ by using a histogram of number of states as function of $p_{\text{fold}}(y_i)$ and $\xi(y_i), i.e., we plot

$$n(p_{\text{fold}}, \xi) = \{y_i | p_{\text{fold}}(y_i) = p_{\text{fold}} \text{ and } \xi_i = \xi\}$$  (15)
as function of $p_{\text{fold}}$ and $\xi$. In the last equation $|A|$ denotes the power of the set $A$ (i.e., its number of elements). As can be seen from Fig. 9 states $y_i$ with higher $\xi(y_i)$ have, generally, a higher folding probability $p_{\text{fold}}(y_i)$. This indicates that $\xi$ (i.e., the number of residues belonging to alpha helices) is indeed a reasonable reaction coordinate. However, one should note that $n(p, \xi)$ is not strictly monotonic in $p$ for given $\xi$. This behavior is mainly caused by the occurrence of isolated helical residues which remain stable for significant amounts of time, cf. the discussion in the preceding section.

In order to compare these analytical results with MC data, we picked 99 random initial configurations (nine for every value of $4 \leq \xi \leq 14$) and then performed $N_{\text{mc}} = 100$ runs starting from these configurations. The folding probability $p_{\text{fold}}^{\text{mc}}(y_i)$ obtained in this way should be exact within 10%, cf. Ref. 7. The folding probability $p_{\text{fold}}^{\text{est}}$ calculated with our method also has a statistical error due to the stochastic nature of the analyzed MC data. Since it is impossible to systematically investigate the scaling behavior of the error of $p_{\text{fold}}^{\text{est}}$ with $N_r$, we simply assume that it is similar to that in the Ising model (cf. the discussion in Sec. III). Then, $p_{\text{fold}}^{\text{est}}$ is also accurate within 10%. Figure 10 shows the comparison of the two datasets. Within the estimated errors the agreement is reasonably good. In fact, within the error bars 72% of the data points lie on the diagonal $y = x$, justifying our assumption that the error of $p_{\text{fold}}^{\text{est}}$ scales as $1/(3N_r^{0.21})$.

However, deviations between both approaches not only occur due to statistical error but also due to the fact that some states are projected out. In order to show that the deviations between the two approaches becomes smaller as $N_{\text{rel}}$ (i.e., the number of states taken into account) increases, we calculated the root-mean-square deviation $\sigma$ of $p_{\text{fold}}^{\text{est}}$ from $p_{\text{fold}}^{\text{mc}}$ as function of $N_{\text{rel}}$. Thus,

$$\sigma^2 = \frac{1}{N_{\text{rel}}} \sum_{y_i} (p_{\text{fold}}^{\text{est}}(y_i) - p_{\text{fold}}^{\text{mc}}(y_i))^2$$

We then found that $\sigma$ is monotonically decreasing with $N_{\text{rel}}$. More precisely we found $\sigma = 0.133, 0.113, 0.097, 0.092$ for $N_{\text{rel}} = 1309, 2120, 3239, 5960$, respectively.

V. DISCUSSION

In summary, in this study we have demonstrated how folding probabilities of microstates can be analytically calculated in systems where the underlying dynamics is described by a master equation. To illustrate our general approach we applied our method to the analysis of MC simulations of the Ising model and a newly developed model for the formation of alpha helices. As these applications show the agreement between the calculated and the precise (respectively, measured) values of $p_{\text{fold}}$ is reasonably good (although not as good as we had hoped for).

As discussed, there are several possible reasons for these deviations. The most important one is the stochastic nature of the analyzed MC data. However, as we have shown our method is much more effective in acquiring a desired accuracy than more conventional approaches: from a few MC simulations the folding probabilities of a whole ensemble of microstates can be calculated simultaneously.

Systematic errors might be caused by the fact that the MC simulations do not reach all microstates of the system. Therefore, the calculated rate matrix does not reflect the full connectivity of phase space. A fact which seems to influence the numerical value of $p_{\text{fold}}$ much more than errors in the estimated transition rates between states which are reached by the trajectories.

Thus, for the described applications a calculation of the folding probability with very high accuracy is nearly impossible. Nevertheless, our analysis shows that our method yields a good estimate for the folding probability which correlates rather strongly with the real values. Our method will be useful for the optimization of coupled simulations especially since only a few MC runs are needed to calculate the folding probabilities for a large number of states. The knowledge of $p_{\text{fold}}$ can be used to classify transient configurations making thus the coupling between simulations more effective: by resetting only to those states which increase $p_{\text{fold}}$ significant speedup in simulation time can be gained.

Future research will concentrate on the study of the properties of the folding probability of microstates in a more general context. From our examples it can be already seen
that $p_{\text{fold}}$ is a reaction coordinate which depends rather strongly on the details of phase space and thus might provide a useful mean to characterize its topology. For example, $p_{\text{fold}}$ depends via the spectrum of the transition matrix crucially on the properties of the free energy (or the Hamiltonian): while for the Ising model all eigenvalues are real, the energy landscape is more complex for the protein folding model, and $|K_{ij}| < 1.0$ is possible. Then, complex eigenvalues $\lambda_i$ can occur which correspond to oscillations between stationary states. It appears to be rewarding to study this phenomenon in more detail and to elaborate on a connection with non-Hermitian localization which is presently studied in a variety of areas of physics.  

### ACKNOWLEDGMENTS

We thank all the members of the Pande group for assistance and useful discussions. This work was supported in part by the ACS-PRF 36028-AC4 grant. B.Z. was supported by an HHMI predoctoral fellowship. P.L. acknowledges support by the Deutsche Forschungsgemeinschaft through the Emmy–Noether program (Le 1214/1–1).

---

11. The fact that the formation of helices can be reasonably well characterized by the reaction coordinate $\xi$ is a consequence from the one-dimensional nature of this structure. However, this approach cannot be easily generalized to the folding of more complex proteins.
12. Again, one should note the big advantage of our approach as compared with the direct calculation of $p_{\text{fold}}$: We are able to achieve the same accuracy by analyzing only 466 runs instead of 9900 runs.
13. This statement is supported by our analysis of a simple MD model for protein folding (data not shown). In these MD simulations, phase space is explored much more thoroughly and the various protein configurations are much more connected. As a consequence, the folding probability of the protein configurations is generally much smaller than in the MC model and only weakly depends on the order parameter.
29. It is worth mentioning that we also analyzed our MC data by using different estimates for $K_{ij}$. For example, for systems in which the mean passage time $\bar{\tau}_i$ of each state $i$ is dominated by a high transition rate to a state $j$, a more appropriate estimate is given by $K_{ij}^\text{est} = p_{ij}/\bar{\tau}_j$. However, we found that our calculated $p_{\text{fold}}^{\text{est}}$ depends only very weakly on the choice of $K_{ij}^\text{est}$ indicating that $p_{\text{fold}}$ depends much more on the connectivity of phase space (which determines the number of nonvanishing elements of $K_{ij}$) than on the precise value of the transition rates.