Kinetics of the coil-to-helix transition on a rough energy landscape

A. Baumketner* and J.-E. Shea†

Department of Chemistry and Biochemistry, University of California, Santa Barbara, California 93106, USA

(Received 11 May 2003; revised manuscript received 26 August 2003; published 3 November 2003)

The kinetics of folding of a fully atomic seven-residue polyalanine peptide in an implicit solvent are studied using molecular dynamics simulations. The use of an implicit solvent is found to dramatically increase the frustration of the energy landscape relative to simulations performed in an explicit solvent [Phys. Rev. Lett. **85**, 2637 (2000)]. While the native state in both implicit and explicit solvent simulations is an α-helix, the kinetics of the coil-to-helix transition differ significantly. In contrast to the explicit solvent simulations, the native state in the implicit solvent simulations is not kinetically accessible at temperatures where it is thermodynamically stable and could not be brought into equilibrium with other conformational states. At temperatures where statistical equilibrium was achieved, the conformational diffusion folding mechanism, found earlier to be adequate for this peptide in an explicit solvent [Phys. Rev. Lett. **85**, 2637 (2000)], is met with only limited success. Issues relating to the evaluation of the quality of implicit solvent models on the basis of thermodynamic criteria only are reexamined.

DOI: 10.1103/PhysRevE.68.051901 PACS number(s): 87.15.He, 87.15.Cc, 87.15.Aa, 87.10.+e

I. INTRODUCTION

In the past decade, molecular dynamics (MD) simulations have become an established and vital tool for the theoretical investigations of proteins [1]. Recent developments in computer hardware technologies have allowed the study of increasingly complex protein systems on modern computers. In numerical studies of atomistic protein models, particularly spectacular success was achieved in the understanding of the thermodynamics of folding [2]. Recent advances in algorithm developments (such as generalized ensemble sampling [3]), as well as the emergence of novel methods to analyze simulation trajectories [4] or even entire classes of protein models [5], have enabled theorists to probe experimentally relevant issues pertaining to the free energy differences between unfolded and native states. Free energy maps constructed as functions of a few selected reaction coordinates can now be routinely generated for small single-domain proteins and peptides. By using special sampling techniques, it also becomes increasingly feasible to discuss folding mechanisms for larger proteins, such as dihydrofolate reductase [2].

In terms of protein folding kinetics, the success of direct computer simulations of fully solvated proteins remains rather limited [6]. Due to the associated computational effort, simulations of folding events on biologically relevant time scales (1 ms or longer) for systems containing thousands of atoms are prohibitively costly. The recent ''brute force'' MD investigation by Duan and Kollman [7] is representative of the complexity of the problem. The failure of a month-long simulation carried out on a CRAY supercomputer to observe a single folding event for a 36-residue peptide clearly demonstrates that systematic numerical studies of protein folding at a microscopic level are still beyond our reach, even on the most powerful parallel computers.

In light of the computational obstacles associated with explicit solvent simulations, the use of implicit solvent models has emerged as a promising alternative for kinetic studies. Over the past decade a large variety of such models have been developed and tested in biological applications [8,9]. The quality of these models is commonly evaluated by judging how well implicit solvent models reproduce low-energy protein structures observed in experiments (or explicit solvent simulations). In the language of energy landscape theory, the quality of an implicit solvent model is evaluated based on how accurately it can predict the location of the native state in the multidimensional conformation space. The location of the lowest-energy state on the energy surface is, however, only one ingredient to understanding protein folding within the energy landscape perspective. Of equal importance is the time scale on which the native state can be reached under physiological conditions, in other words, the kinetic accessibility of the native state. The issue of kinetic accessibility is intrinsically tied to the degree of frustration in a protein model, or the relative roughness of its energy landscape. The latter, by nature an abstract qualitative concept, can be quantified, among other means, by the thermodynamic ratio $R=T_f/T_g$ of the folding temperature T_f to the glass transition temperature T_g [10]. The relevance of this ratio for theoretical modeling of proteins lies in its ability to categorize protein models as fast or slow folders. The native state of a protein with $R<1$ is not accessible at physiologically relevant temperatures, despite having the lowest energy of all possible conformations. Such a system is frustrated and is considered a poor, slow-folding protein model. Protein models with $R>1$, on the other hand, are believed to fold rapidly and may offer a more accurate representation of the folding properties of real proteins.

In this paper, we demonstrate that in developing implicit solvent models, the kinetic accessibility of the native state should be given as much consideration as the location of the thermodynamic ground state. Using all-atom simulations of a

*Permanent address: Institute for Condensed Matter Physics, 1 Svientsitsky Street, Lviv 79011, Ukraine. Email address: andrij@icmp.lviv.ua

†Author to whom correspondence should be addressed. Email address: shea@chem.ucsb.edu

seven-residue alanine peptide, we show that despite folding to the native state predicted from simulations in an explicit solvent [11], an oversimplified implicit solvent model produces unrealistic degrees of frustration in the underlying potential energy surface. At temperatures where the native state is expected to be stable, simulations were dominated by long traps in misfolded conformations and failed to reach equilibrium over the time period of 6μs, encompassing thousands of folding events. Equilibrium was only achieved at higher temperatures where the statistical weight of the native state is less than 1%. Even at these temperatures, important aspects of the kinetics observed with our implicit solvent model disagree with those found in previous explicit-solvent simulations. In particular, the kinetic mechanism of conformational diffusion search, found to be adequate for this peptide in explicit solvent simulations [11], is met here with only limited success. The theoretical mean first-passage time evaluated by the diffusion-equation formula of Bryngelson and Wolynes [12,13] is about five times longer than the folding time estimated directly from simulations. Poor agreement between earlier results [11], obtained in explicit solvent conditions, and those obtained here for the same peptide in implicit solvent, highlights the need to consider both kinetic and thermodynamic criteria when assessing implicit solvent models (for example, through the calculation of long-time correlation functions [14]) rather than relying solely on reproducing the native state [9,15].

The paper is organized as follows. In Sec. II we describe the simulations methods as well as the all-atom model of the polyalanine peptide. The folding kinetics of the model peptide are discussed in Sec. III. Conclusions are presented in Sec. IV.

II. MODEL AND COMPUTATIONAL DETAILS

The structure, thermodynamics, and dynamics of alanine polypeptides have long been a subject of great interest in protein science. The high helical propensity of alanine-based homopolymers makes these peptides ideal candidates to test various theories of protein folding, in particular theories of the helix-coil transition [16]. Over the past decades, formation of helices in polyalanines and their stability have been the subject of intense experimental studies [17]. Theoretical approaches to this problem have involved a wide range of methods and models [11,18–22]. Atomistic models of polyalanines, solvated in explicit water, were studied by molecular dynamics [11] and peptide growth methods [19] to elucidate the details of the coil-helix transition. Klein *et al.* [21] used simulated annealing methods with a solvent accessible surface area implicit solvent model to locate the native state of polyalanines and study their conformational statistics at finite temperatures. An EEF1 implicit solvent model was used by Levy *et al.* [22] to investigate the effects of solvation on free energy surfaces (which in turn reflect peptide conformational preferences for α-helical or β-sheet conformations). The treatment of long-time dynamics in polyalanine peptides, barely tractable in regular molecular dynamics simulations, was addressed by Huo *et al.* [20] using the reaction path method.

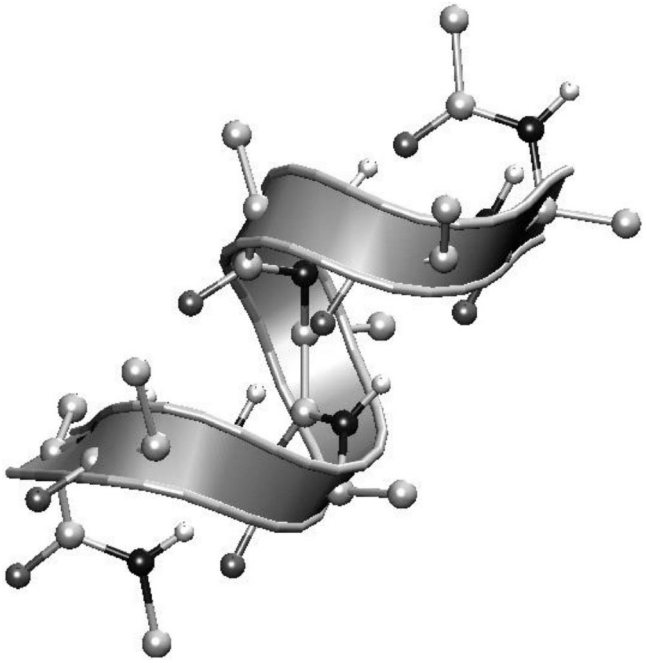

FIG. 1. Native α-helical state of the polyalanine model considered in the present study.

A. Polyalanine model

In this paper we consider a seven-residue long alanine polypeptide Ace-$(Ala)_7$-Nme. The CHARMM PARAM19 force field is used for interatomic interactions with all heavy atoms and hydrogens bound to nitrogen atoms considered explicitly. No cutoff distances are applied for long-range van der Waals and Coulomb interactions. To model the presence of solvent, we used a distance-dependent dielectric constant $\epsilon = r$ [9] in our simulations. It should be noted that setting $\epsilon = r$ provides a very crude way to account for the electrostatic part of the solvation energy. Nevertheless, simulations employing this simplified implicit solvent have shown some encouraging results [9] with regard to the nature of the native state and distance-dependent dielectric constants are widely used nowadays in biological applications [23,24]. Simulations using a distance-dependent dielectric constant do not require more computational effort than simulations performed in vacuum and in this respect compare favorably with other implicit solvation models, such as those based on the surface accessible area calculations [8]. Since our aim is a detailed study of the folding kinetics in implicit solvent conditions, it is imperative to keep the computational burden of the simulations as low as possible.

To determine the native state of the considered polyalanine we launched a series of simulated annealing runs. The computations were organized according to the logarithmic temperature protocol $T_{i+1} = \alpha T_i$ where T_{i+1} and T_i are temperatures at two consecutive cooling stages. By setting the parameter α in this protocol to 0.9 the temperature was reduced from 800 K to 10 K, allowing 10^5 equilibration time steps at each particular temperature. A total of 50 simulations were launched, of which 17 reached the native basin (in our

TABLE I. The backbone dihedrals (ϕ,ψ) for the studied polyalanine model in the α-helical native state conformation.

Residue no.	1	2	3	4	5	6	7
ϕ,ψ	$(-62,-42)$	$(-63,-32)$	$(-74,-39)$	$(-62,-38)$	$(-65,-39)$	$(-70,-34)$	$(-68,-38)$

definition, conformations with rms deviations from the native state of less than 0.6 Å).

We found that the native state of the studied peptide model is an α helix, as shown in Fig. 1. Characteristic α helical patterns of hydrogen bonds, formed between hydrogens of residue i and oxygens of residue $i+4$, are readily seen in the figure. The energy of the native helix is -94.6 kCal/mol. The values of the seven backbone (ϕ,ψ) dihedral angles of the native state compiled in Table I are consistent with those obtained in previous simulations [11,21]. We note that recent experimental studies indicate that polyalanine sequences with fewer than ten residues may preferentially adopt polyproline II rather than α-helical structures [18]. The higher helical propensity found in simulations may result from small errors in the force fields [25].

B. Details of molecular dynamics simulations

The CHARMM macromolecular modeling program [26] was used to perform the molecular dynamics simulations. Six canonical ensemble simulations at temperatures $T=470$, 500, 550, 600, 700, and 800 K were carried out using the velocity Verlet algorithm and the Nosé-Hoover [27] thermostat method with the q friction parameter set to 50 ps. Throughout the work we employed the integration time step $\delta t=1$ fs which was chosen such that the conformational statistics of the model generated for a set of temperatures did not change noticeably when shorter integration time steps were used. The length of the hydrogen-nitrogen bonds (the only bonds involving hydrogen atoms in the present system) was kept constant during the simulations through the SHAKE algorithm [28]. The lengths of the simulations depended on the temperature considered. The longest total simulation time was 7 μs, at the lowest temperature considered, $T=470$ K. At the highest temperature $T=800$ K the total physical time reached by the simulations was 0.15 μs. During the simulations, peptide conformations were stored into a file at a regular time interval of $\Delta t=400$ time steps. Special care was taken in selecting this value of the time interval. Since our aim is to generate folding events, the time resolution with which conformations are saved along the trajectory should be no longer than the average time the protein spends in the native state, here the time required to reach conformations differing from the native state by more than 0.6 Å in rms. Otherwise some of the folding events may pass unregistered in the simulations, thereby distorting the final results for the folding time. To avoid such situations we verified explicitly, by running short unfolding simulations, that the model stays in the native state longer than 400 time steps on average, at all temperatures considered.

III. RESULTS

A. Slow kinetics of folding

Over the course of the simulations the peptide underwent multiple rounds of folding and unfolding. The rms displacement of a given conformation from the native state was taken to be the progress variable of folding. Conformations with rms <0.6 Å were considered to be folded. For the unfolded state there is no unique way to define its boundaries, since, technically, every non-native conformation is unfolded. For the sake of comparison with theoretical calculations that follow later in the text, we consider here conformations with $3.7>$ rms >3.6 Å to be unfolded. A justification for this choice, as well as a discussion of the nature of the unfolded state in determining folding rates will be given in a later section. By analyzing the simulation trajectories, sequences of folding (as well as unfolding) times were generated at the six temperatures considered. To give an idea of the extent of data considered, we note that the sequence of folding times recorded for $T=500$ K contained 7582 entries. Comparable numbers were generated at each temperature. The sequences were used to construct distributions of the folding time $P_f(\tau)$ as well as the probability distribution $P(\tau)$ for the protein to remain unfolded at time τ, provided that folding started at time zero (i.e., the survival probability). The latter is shown in Fig. 2(a) for a range of temperatures. The trend for folding dynamics to slow down at low temperatures is immediately apparent in the figure. At the highest simulated temperature, 800 K, the survival probability $P(\tau)$ is determined by a single time scale. At this temperature, a single-exponential function $e^{-\alpha\tau}$ can successfully fit the simulation data. As the temperature decreases, the survival probability begins to spread. Events occurring at longer times become increasingly relevant at low T and the relaxation processes in the system begin to split into two different time scales. This feature already appears at $T=700$ K and becomes increasingly pronounced at lower temperatures where the survival probability acquires slowly relaxing tails and becomes strongly nonexponential. Our attempts to fit $P(\tau)$ at low temperatures to the biexponential form $\phi e^{-\alpha\tau}+(1-\phi)e^{-\beta\tau}$, characteristic for folding mechanism that splits folding pathways into slow and fast channels, were unsuccessful. We also failed to fit the long-time tails of $P(\tau)$ either to a single or stretched exponential function $fe^{-(\alpha\tau)^{\beta}}$. The nature of the exact functional form of $P(\tau)$ remains ambiguous. Although the low-temperature kinetics could be represented as a sum of more than two exponentials, it is unclear how many exponents are required and, more importantly, what is the physical meaning of these exponents.

As the temperature is lowered from 800 to 470 K, the time scale of the relevant relaxation processes varies wildly,

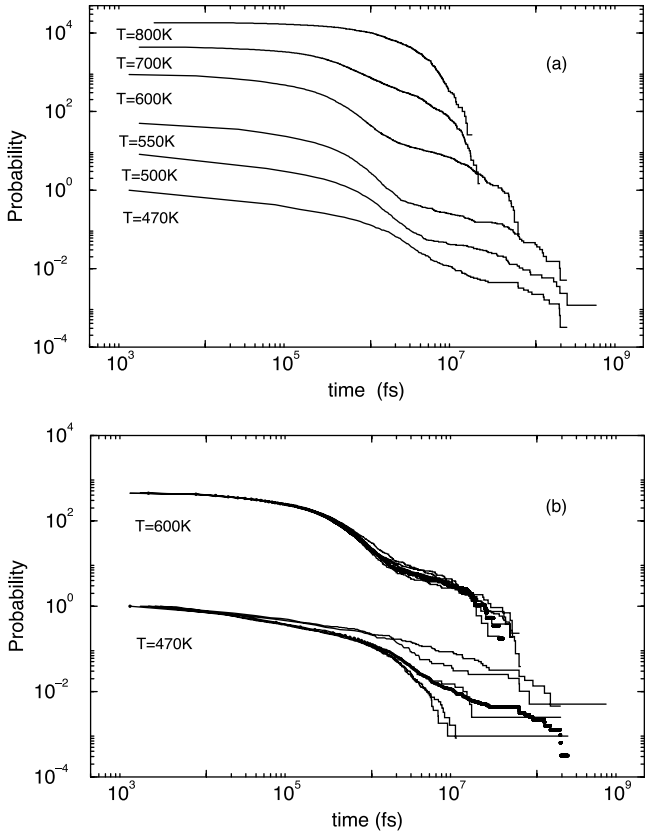

FIG. 2. Probability distribution $P(\tau)$ of the studied polyalanine model to remain in an unfolded state at time τ when folding is initiated at time zero. It is assumed that upon folding the protein cannot escape from the native state. (a) $P(\tau)$ at varying temperature, (b) $P(\tau)$ at two selected temperatures demonstrating the extent of statistical convergence of the data. The curves in both parts are shifted upward for visual convenience.

almost over two orders of magnitude [Fig. 2(a)]. In dealing with slow processes of this nature, the statistics of the simulation data become a key issue and it is critical to ensure that statistical convergence has occurred in the properties of interest. We test the statistics by splitting the trajectories into five nonoverlapping parts and computing the distribution $P(\tau)$ for each of them separately. This allows us to locate the portions of $P(\tau)$ with poor convergence as those displaying significant scatter in different parts of the trajectory. The results of this analysis are shown in Fig. 2(b) for two temperatures, 470 K and 600 K. It is clear from the figure that at the higher temperature the survival probability is sufficiently converged at all times. At the lower temperature, on the other hand, different parts of the same trajectory produce different values for $P(\tau)$. The same results were obtained for the second lowest temperature studied in this paper, 500 K. As illustrated in Fig. 2(b), with the exception of the short time behavior ($\tau<2$ ns), the survival probability cannot be reliably determined from the simulation data obtained in the present work at both $T=470$ and 500 K. In particular, it is impossible to reach a definite conclusion regarding the buildup of the linear segment at long times in Fig. 2(a) as the temperature decreases.

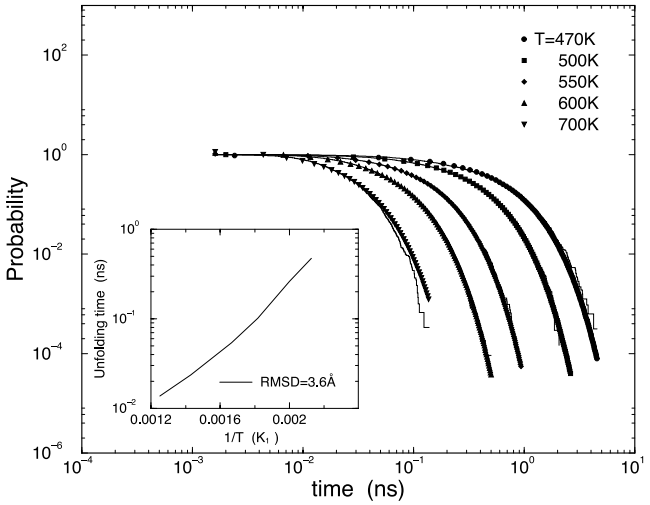

FIG. 3. Survival probability for the folded state at different temperatures. The solid line corresponds to the simulation data, while the symbols represent the fit of the data to an exponential function. The unfolding time is plotted as a function of inverse temperature in the inset. Conformations reaching an rms deviation of 3.6 Å from the native state are considered to be unfolded.

It is interesting to note that the kinetics of unfolding (i.e., of the helix-to-coil transition), differ significantly from the kinetics of folding. The survival probability function for the folded state for a range of temperatures is plotted in Fig. 3. The kinetics are exponential at all temperatures considered, although the fit becomes poorer at short times as the temperature is increased. The unfolding time as a function of inverse temperature is plotted in the inset of Fig. 3. The unfolding time increases with decreasing temperature and the kinetics appear to be Arrhenius at low temperatures.

The linear segment in the distribution of the first-passage times could indicate a power-law dependence of the survival probability on time. This type of probability distribution, the so-called Lévy distribution, was theoretically predicted for proteins on the basis of the random energy model [29–31]. If true in the case of more realistic protein models, this finding will have profound consequences for the protein folding problem, especially for the stability of the native state at low temperatures. We illustrate our point with the following example. Let us assume that the folding time τ satisfies a Lévy power-law distribution:

$$\overline{P}_f(\tau) \sim \frac{\mu \tau_b^{\mu}}{\tau^{1+\mu}}, \quad \tau \to \infty, \tag{1}$$

where the exponent $\mu<1$ and τ_b specifies the time scale of the problem. For such statistical distributions, which we refer to as broad to distinguish them from the normal narrow distributions, the mean value of the relevant stochastic variable, in our case the folding time, does not exist. The folding time in this situation is characterized not by its mean value, but rather by its most probable value which is accurate up to an order of magnitude and not reproducible in repetitive experiments. Let us further assume, as suggested by the present results, that the unfolding time obeys a narrow distribution:

$$\overline{P}_u(\tau) \sim \frac{1}{\tau_u} e^{-\tau/\tau_u}, \quad \tau \to \infty, \tag{2}$$

with well-defined finite mean value τ_u (which nevertheless can be fairly large at low temperatures below the melting point). For refolding experiments carried out for a single protein that exhibits the above patterns of folding and unfolding statistics, the time average of the probability P_N to be found in the native state is simply given by the ratio of the time spent in the native state T_N to the total observation time $T_N + T_U$:

$$P_N = \frac{T_N}{T_N + T_U}, \tag{3}$$

where T_U is the time spent by the protein in the unfolded state. For a large number N of folding/unfolding events occurring over a long observation time Θ, the folding statistics [Eq. (1)] implies that [32] $T_U = \xi \tau_b N^{1/\mu}$, where ξ is a random variable of order 1. Likewise, the unfolding kinetics [Eq. (2)] predicts that the time spent by the protein in its folded state is $T_N = \tau_u N + O(\sqrt{N})$. Substituting T_N and T_U into Eq. (3) one finds that in the leading order in N the probability of the native state is

$$P_N = \frac{\tau_u N}{\tau_u N + \xi \tau_b N^{1/\mu}} \sim \frac{\tau_u}{\xi \tau_b} N^{(\mu-1)/\mu} \sim \frac{1}{\Theta^{1-\mu}} \to 0, \quad \Theta \to \infty. \tag{4}$$

In this expression we made use of the approximation for the observation time $\Theta \sim T_U$ appropriate for stochastic processes governed by Lévy distributions [32]. Equation (4) predicts that the probability of the native state population vanishes at long Θ when the exponent μ of the folding time distribution $\overline{P}_f(\tau)$ is lower than 1. In other words, due to the fact that folding times are infinitely longer that unfolding times, an isolated protein obeying the Lévy folding distribution [Eq. (1)] will spontaneously unfold if the experiment is allowed to continue over a sufficiently long observation time. A similar conclusion that proteins should denature when observed at sufficiently long times can also be reached for an ensemble of proteins [32]. The specific exponent characterizing the time decay of the ensemble average of the native state population is different, however, from the one given above for the time average. This disagreement is a direct consequence of ergodicity breaking in systems subject to Lévy type statistics. In such systems, time averages are not equal to ensemble averages and the rules of usual thermodynamics cease to apply. This holds true of the requirements for thermodynamic stability of the native state, as can be seen from the above discussion.

B. Conformational diffusion mechanism of folding

According to the energy landscape theory [10,33,34], folding of a protein can be depicted as stochastic motion on the statistically averaged free energy surface defined in terms of a few collective variables, or reaction coordinates. When the space of the reaction coordinates is taken to be one-dimensional, characterized by a variable χ, and the stochastic motion of the protein on the energy surface is considered to be Brownian, such that χ satisfies the Smoluchowski diffusion equation, it is possible to derive an explicit expression for the mean first-passage time for folding [12]:

$$\tau_f = \frac{1}{D} \int_{\chi_{unf}}^{\chi_{fol}} dx \int_{\chi_{ref}}^{x} dy e^{\beta[U(x)-U(y)]}, \tag{5}$$

where $U(\chi)$ is the free energy profile in χ, β is the inverse temperature, and D is the conformational diffusion constant which here is taken to be independent of χ. The integration in Eq. (5) is carried from the states corresponding to the unfolded ensemble χ_{unf} to the native states χ_{fol}. In the diffusion-equation terminology χ_{unf} reflects the initial conditions of the equation, χ_{fol} is an absorbing boundary (the protein cannot unfold once it has folded) and χ_{ref} denotes a reflective boundary of the equation. The folding process of a protein within this theoretical formalism is fully defined by two terms: the free energy surface $U(\chi)$ and the conformational diffusion constant D. The free energy profile $U(\chi) = -k_B T \ln[P(\chi)]$ along the rms reaction coordinate as well as the related probability distribution function $P(\chi)$ are shown in Fig. 4 for the studied temperatures. A detailed discussion on the probability distribution will follow later in the text.

Recently, the validity of the conformational diffusion formalism as implemented on the basis of Eq. (5), was tested [11] in explicit solvent simulations on a five-residue alanine polypeptide by using AMBER94 force field. By computing the free energy surfaces directly from simulations and obtaining a value of the conformational diffusion constant from fits to the theoretical and simulation mean first-passage times (MFPT), the authors concluded that the conformational diffusion search adequately describes the folding mechanism of the polyalanine peptide. This result was supported by the observation of nonexponential kinetics in the model, which arise naturally within the conformational diffusion theory. Indeed, according to this theory, transitions of a protein into the native state are considered from a multitude of nonnative conformations that contribute to the folding time with broadly distributed rates.

More recently, it was shown that the conformational diffusion coefficient D can also be evaluated directly in simulations without resorting to potentially unreliable fitting procedures [13]. Specifically, D can be estimated from the free energy profile and time course of the reaction coordinate χ as the integral over $\psi(t)$, the time correlation function of χ, and the generalized force exerted on this variable $F[\chi] = -\partial U(\chi)/\partial \chi$. The utility of this approach was demonstrated by calculating the MFPT of a small β sheet off-lattice protein. Here we employ the formalism of Ref. [13] to evaluate the folding time of the studied polyalanine, using the RMS deviation as the reaction coordinate. The theoretically calculated MFPT as well as the one obtained directly from our simulations are shown in Fig. 5(a). The MFPT curve has the characteristic V-shaped structure reported in earlier simulations [13] for the temperature dependence of protein folding times. The increase in MFPT at high temperatures arises from the low statistical probability of the native state, while

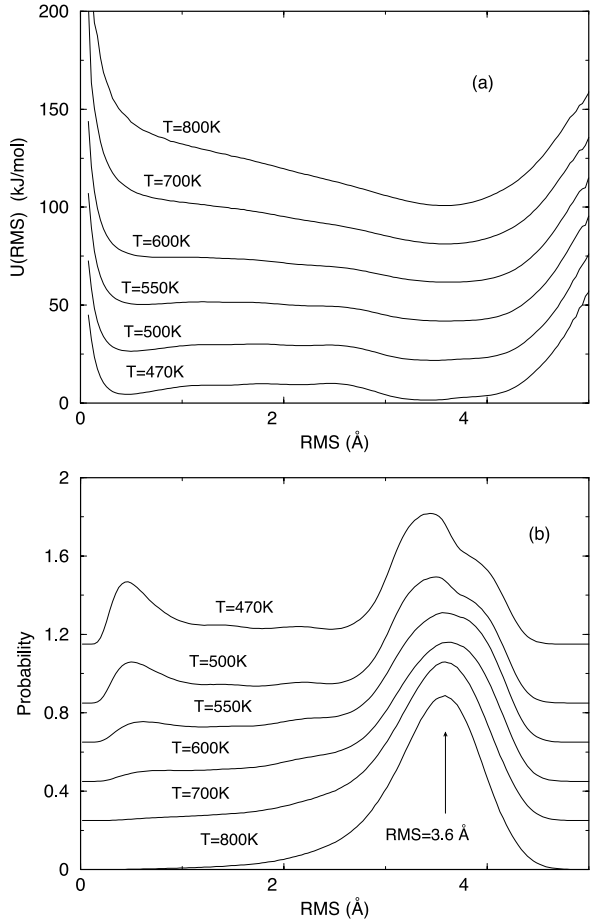

FIG. 4. (a) Free energy profiles along the rms deviation from the native state. (b) Probability distribution functions of rms computed for the seven-residue polyalanine model studied in the present work. For visual convenience the curves in both parts have been shifted upward.

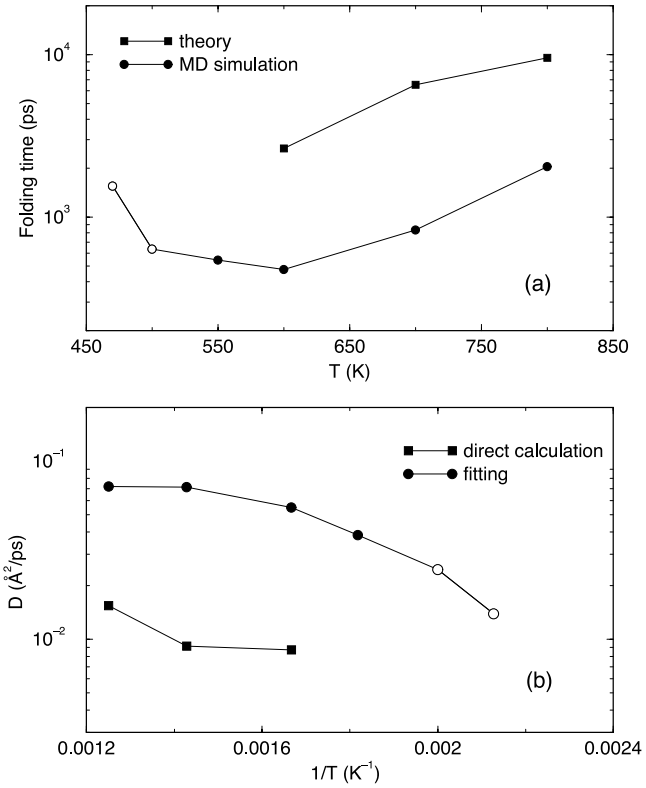

FIG. 5. (a) Mean first-passage time obtained for the studied polyalanine model through the diffusion-equation formula (5) and that calculated directly in folding simulations. (b) Conformational diffusion coefficient D computed directly from simulation trajectory and that obtained by fitting the theoretical MFPT (5) to the one obtained in folding simulations. In the folding simulations the unfolded state ensemble comprised conformations with rms$=3.6$ Å. Open symbols refer to the statistically unreliable data.

the increase at low temperatures is a consequence of kinetic trapping. Note that in the explicit solvent simulations [11] only the lower temperature part of the MFPT was reported. As mentioned above, our simulations failed to reach equilibrium for the two lowest temperatures considered. To reflect this fact, the data for τ_f at these temperatures are marked by open symbols in Fig. 5(a). Also for these temperatures, and additionally for $T=550$ K, we failed to obtain convergent results for the time correlation function $\psi(t)$. As a consequence, the theoretical data for τ_f at these temperatures are missing in the figure. The equilibration was successful at higher temperatures where it affords an assessment of the validity of the conformational diffusion folding mechanism for the present model. As can be seen from Fig. 5(a), there are significant discrepancies between the theoretical results calculated using formula (5) and those obtained directly from simulations. In simulations, a conformation was considered to be folded when its rms deviation from the native state was less than 0.6 Å. Simulation and diffusion-equation folding times agree well qualitatively, both predicting a decrease in folding time as the temperature is lowered. The quantitative difference between the two methods, however, is fairly large: the theory predicts MFPT at least five times longer than that obtained from simulations. This discrepancy should be compared to a difference of less than twofold obtained in two previous applications of the conformational diffusion formalism [13,35] for different minimal protein models. With these points considered, we must conclude that the success of formula (5) in predicting the folding time for the present model is rather limited.

When assumed to be correct for the rms deviation from the native state, the formalism of Brownian dynamics (or equivalently the Smoluchowski equation [11]) can be used to evaluate conformational diffusion coefficient D. The diffusion coefficient represents clear theoretical interest as one of the ingredients involved in formula (5) that allows direct evaluation of the folding time once the free energy profile is known. It can be calculated by matching theoretically computed MFPT with that obtained from the simulation trajectory directly. Any disagreements between the diffusion coefficient computed in the direct way and that obtained by the fitting procedure signify deviations of the dynamics of the reaction coordinate (rms) from the Brownian dynamics. Diffusion coefficient D computed directly in simulations and that obtained through fitting of the MFPTs calculated by using conformations with rms$=3.6$ Å as the unfolded en-

semble, are plotted in Fig. 5(b) as a function of inverse temperature $1/T$. As expected from Fig. 5(a), the two diffusion coefficients disagree rather strongly at temperatures where the direct evaluation of D is possible. Except for the qualitative trend to increase with the temperature the diffusion coefficients computed by the two different methods have little in common. As mentioned before, this disagreement provides clear evidence of the failure of the present theoretical formalism to describe the dynamics of the reaction coordinate. Another interesting finding following from Fig. 5(b) is that the temperature dependence of the diffusion coefficient D does not obey the Arrhenius law at low temperatures as predicted theoretically [36] and in computer simulations [11]. It should be noted, however, that our data at low temperatures are not sufficiently reliable to allow any quantitative conclusions.

In addition to comparing the folding times from simulations with those from the diffusion-equation formula (5), we can also assess the validity of the conformation diffusion approach for our model by investigating the influence of the initial folding conditions on folding time. More specifically, we seek to determine how this influence is captured in Eq. (5). This question brings up the more general and often overlooked matter of the role of the unfolded state ensemble in protein folding. In the following section we discuss the influence of the nature of initial folding conformations on the folding time of our peptide.

C. The role of the unfolded state ensemble

We briefly mentioned in an earlier section that, in contrast to the native state, the unfolded state does not consist of a single conformation, or even of a few conformations. Strictly speaking, every non-native conformation belongs to the unfolded state and a statistical approach to defining the unfolded state as an ensemble of conformations, or a distribution, is hence necessary. Parameters of such distributions are expected to depend strongly on the conditions of the protein environment such as the temperature or chemical composition of the solvent. These conditions determine structural and dynamical properties of the unfolded, or denatured, state ensemble. Since the structure of the unfolded state can vary with denaturant conditions, it is reasonable to assume that the kinetics of folding will also be affected by the degree of denaturation.

The nature of the unfolded state has been a long-standing, although at times underappreciated, area of research in protein folding [37]. A number of important developments in recent years have attracted a renewed interest in this problem. In particular, it was found that, contrary to well-established views, conformations of the denatured state are not random coils but may possess a significant amount of secondary structure [38]. The size of the unfolded conformations, as estimated by the radius of gyration, also appears to be substantially smaller than would arise from completely structureless states [38,39]. The structure of the unfolded state ensemble has been shown to strongly depend on the denaturing conditions. For instance, denaturation by different chemical means produces unfolded ensembles of similar size, but which are quite different in conformational nature [39]. From a dynamical point of view, proteins in the unfolded state do not display random coil dynamics [40].

These experimental findings regarding the structure and dynamics of unfolded proteins emphasize the need for a theoretical investigation of how the nature of the unfolded state affects protein folding. Such an investigation may provide useful insights into problems where the denatured state plays a critical role. For example, one of the theories of chaperone-assisted folding [41] asserts that the protein is mechanically unfolded upon binding to the the chaperonin cavity [42]. The nature of the unfolded state produced through this mechanical stretching is not known. It is also not clear how this unfolded ensemble will affect protein folding times in comparison to unfolded ensembles generated by other means.

Questions pertaining to the structure of denatured states and its influence on folding kinetics are most conveniently addressed using computer simulations. The large number of non-native conformations generated in the simulations of the present work allows us to probe the effects of the unfolded state on the folding time of the model. The probability distribution of rms calculated at different temperatures is shown in Fig. 4(b). Recall from our earlier discussion that formula (5) was derived by considering that the state of unfolded conformations can be described by a single parameter $\chi = \mathrm{rms}$. With the protein unfolded at high temperatures, it is apparent from the figure that the unfolded state cannot be assigned a single value of rms. A rather broad distribution $P_i(\chi)$ over different values of rms characterizes the thermally denatured ensemble. Accordingly, folding time of simulations initiated from this type of unfolded ensemble is given by the expression

$$\tau_f^{th} = \int_{\chi_f}^{\infty} P_i(\chi)\,\tau(\chi)\,d\chi, \tag{6}$$

where $\tau(\chi)$ is the folding time for simulations started from conformations with $\mathrm{rms}=\chi$. We note that previously $\tau(\chi)$ was taken to represent the actual folding time of the protein when χ was set to χ_{unf}. The folding time $\tau(\chi)$ can easily be calculated in simulations when a sufficient number of trajectories started from unfolded conformations are available. It can also be estimated within the conformational diffusion approximation through formula (5). The dependence on the unfolded state ensemble enters Eq. (6) through both $P_i(\chi)$ and $\tau(\chi)$. Below we investigate $\tau(\chi)$ computed in the present simulations for five different values of the initial conditions $\chi_i < \mathrm{rms} < \chi_i + \Delta\chi$, $i = 1,5$. The parameter $\Delta\chi$ was chosen such that the ensemble of the initial folding conformations is statistically meaningful. It was found that a value $\Delta\chi = 0.2$ is large enough to select at least 1000 entries from the available pool of unfolded states for all χ_i investigated at all temperatures. The initial values of the rms deviation were taken to be $\chi_i = 2.4$, 2.8, 3.2, 3.6, 4.0 Å. As seen from Fig. 4(b), conformations with $\mathrm{rms} = 3.6$ Å enter the thermally denatured ensemble with highest probability at all temperatures. This is why this value was chosen in the preceding section for the quantitative comparison of the folding times computed using theoretical and simulation methods. In Fig. 6

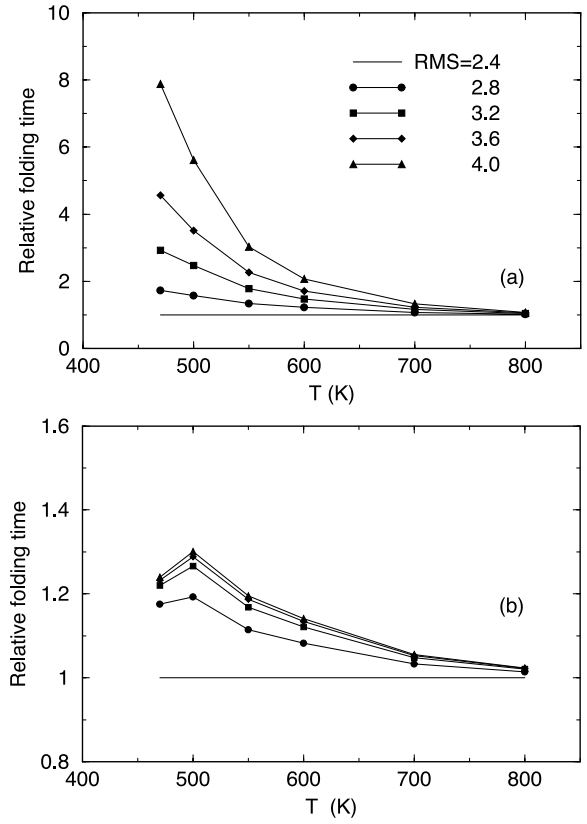

FIG. 6. Relative folding time $\tau(\chi)/\tau$ ($\chi=2.4$ Å) computed for the polyalanine model studied in the present work as a function of temperature. (a) Results of direct simulations, (b) data generated by the diffusion-equation formula (5).

we display the ratio of folding time calculated for different χ to the folding time for the smallest value of rms$=2.4$ Å. The ratio $\tau(\chi)/\tau(\chi=2.4\ \text{Å})$, plotted as a function of temperature, is computed directly from the simulation trajectory in Fig. 6(a), as well as using the diffusion-equation formula in Fig. 6(b). Figure 6(a) clearly shows that the folding time increases as the unfolded states depart from the native state (as reflected by larger values of rms). An indirect confirmation of this trend can be found in a paper by Fersht *et al.* [43] which reports folding times as a function of initial temperature and *p*H levels. The authors report an increase in folding time as the temperature is raised, which is in good qualitative agreement with our simulations. Regrettably, the data collected in the present work are not sufficient to compute $\tau(\chi)$ on a dense enough mesh to allow full quantitative predictions of the folding time as a function of initial temperature.

We mentioned previously that in order to assess the applicability of Eq. (5) to our system, it is not sufficient to merely consider a single value of rms for the initial folding conformations χ_i. Rather, it is necessary to test how this formula works when χ_i is varied. A similar observation that initial folding conditions should be varied in order to validate the conformational diffusion folding mechanism was made in the recent experimental paper by Gai *et al.* [44]. Figure 6(b) shows the variation in folding time, predicted by formula (5), as the rms of the initial folding conformations is increased. We note that since D is taken to be independent of the reaction coordinate χ, the conformational diffusion constant drops out in folding time ratios. All the changes observed in Fig. 6(b) for $\tau(\chi)$ hence originate purely from the free energy surface contributions. As is apparent from Fig. 6(b), the qualitative trend of the theoretical folding time to increase as the initial ensemble departs from the native state agrees with the results for $\tau(\chi)$ obtained from direct simulations. A quantitative comparison of Figs. 6(b) and 6(a) clearly shows, however, that the two folding times are off by almost an order of magnitude. In agreement with our earlier observations, we conclude that on the basis of the data shown in Fig. 6, the conformational diffusion folding mechanism, as implemented through formula (5), has only limited success in describing the folding kinetics of our polyalanine model.

IV. CONCLUSIONS AND DISCUSSION

We find evidence of complex, strongly nonexponential kinetics in molecular dynamics simulations of an atomistic model of a seven-residue polyalanine peptide in an implicit solvent. As the temperature of the peptide is lowered, the distribution of folding time changes from the narrow distribution characteristic of fast relaxation processes, to a broad one, reflective of slow relaxation with long-time tails. Folding at high and low temperatures takes place on disparate time scales, with the relaxational statistics of the model becoming increasingly dominated at low temperatures by single events of trapping in misfolded conformations.

This onset of slow dynamics renders the quantitative characterization of kinetic (such as mean folding time) as well as thermodynamic properties (for instance, distribution of rms) highly problematic in computer simulations. We illustrate this point for our low-temperature simulations for which we failed to achieve full equilibration. The simulations carried out at two lowest temperatures $T=470$ K and 500 K covered physical time of 7 μs, i.e., more than 10^4 typical relaxation times. We consider a typical relaxation time to be on the order of 500 ps, the value of the folding time at $T=550$ K (the lowest temperature for which reliable data were obtained). Under conditions where the underlying stochastic process obeys the laws of normal distribution, averaging over 10^4 independent events should be sufficient to ensure statistical convergence of the data. Clearly, the failure to equilibrate the system may be of technical nature, due to an insufficient number of integration steps performed in the simulations. But it may also be of conceptual origin, indicating the presence of unusual statistical laws governing the folding kinetics of the present model. For example, if the folding time of the model obeys broad Lévy statistics [32] with a power-law exponent smaller than 1, the system becomes nonergodic and convergence in time averages cannot be achieved no matter how long the simulations are run. Unfortunately, it is impossible to determine on the basis of the available data whether this scenario is actually the case in the studied model. Additional simulations are required to clarify this issue.

Another important trait of the folding kinetics of the present model concerns the temperature at which slow dynamics sets in. It is apparent from Figs. 2(a) and 4(b) that at

temperatures where $P(\tau)$ becomes broad, the statistical weight of the native state is negligibly small (less than 1% at $T=600$ K). In other words, the folding reaction is slow at temperatures for which the native state has an occupation probability greater than 0.5, i.e., becomes thermodynamically stable. This folding scenario in which the protein lacks fast kinetics pathways linking the unfolded state conformations to the native one at physiologically relevant temperatures has been predicted [36] on the basis of spin glass theories. This was later confirmed in simulations using lattice [45] as well as off-lattice protein models [46], in which the slow pathways become preferentially populated when the temperature drops. Within this scheme, glassy kinetics arise from relative ruggedness, or frustration, of the free energy surface associated with the protein. The amount of frustration in a protein is quantified by the thermodynamic ratio of the folding temperature T_f to the glass transition temperature T_g. Depending on the numerical value of the ratio $R=T_f/T_g$ there are two possibilities for folding efficiency. If $R>1$ folding proceeds rapidly and displays predominantly exponential kinetics. Slow folders, on the other hand, have ratios $R<1$ and are usually characterized by nonexponential folding kinetics. Considering this classification of proteins, we argue that the model studied in this work belongs to the class of slow folders. Accordingly, we associate the strongly nonexponential and slow kinetics observed in our study with significant levels of frustration in the model.

The kinetic behavior observed here for polyalanine in an implicit solvent should be compared to the results of Ref. [11] in which a similar peptide was studied in an explicit solvent. The folding time of 100 ps obtained in an explicit solvent is comparable to 500 ps obtained here. In addition, nonexponential kinetics were also observed in explicit solvent conditions. A significant difference, however, lies in the fact that simulations in an explicit solvent were able to reach equilibrium at physiologically relevant temperatures of 300 K over much shorter time scale (10 ns) [11] than in our implicit solvent simulations (7 μs). The discrepancy between the kinetic patterns observed for essentially the same protein model but in different solvents implies that the kinetics of folding is a major factor to consider when assessing or developing implicit solvent models. Indeed, as the present model demonstrates, stating that an implicit solvent generates low-energy structures similar to those observed in experiments or an explicit solvent [9] is not sufficient to assess the usefulness of the model for dynamical simulations, as the underlying potential energy surface may possess excessive amounts of frustration. In this respect, an approach adopted by Freed *et al.* [14], in which the aspects of conformational dynamics are tested by calculating time correlation functions in implicit solvents and making comparisons to those obtained in explicit water simulations, seems to be more appropriate.

An additional fact that lends support to the idea that both thermodynamic and kinetic criteria should be taken into account in designing implicit solvent models is that our simulations carried out in an implicit solvent reach conflicting conclusions from those performed with explicit water molecules [11] in terms of the applicability of a conformational diffusion folding mechanism. The conformational diffusion search envisions the folding reaction, which takes place in the multidimensional conformational space, as a diffusive motion along some appropriately chosen progress variable. Depending on whether or not a significant free energy barrier must be overcome on the way to the native state, the folding may be two state or multistate. Recent simulation in an explicit solvent [11] has shown that if the configurational diffusion constant D, which characterizes how freely a protein can interchange conformations, is independent of the progress variable, the resulting theoretical folding formalism can be used to study helix formation in alanine polypeptides. We applied this same formalism in our study to a polyalanine peptide in an implicit solvent and found that its success is limited. Not only do we obtain large fivefold discrepancies in the folding times computed theoretically and directly from the simulation trajectory, but the behavior of the folding time with respect to variations in the unfolded state ensemble is not satisfactorily captured by the diffusion-equation formula (5). It can be stated with certainty that due to the difference in levels of frustration between the implicit and explicit solvent models the diffusive-dynamics formula (5) in its present form does not describe the folding kinetics of our peptide model. Clearly, one would need to consider higher-level models, for instance one in which the diffusion constant depends on the reaction coordinate [29], or a multidimensional reaction coordinate [47], to bring theoretical and computer simulation results in better agreement.

ACKNOWLEDGMENTS

We thank Angel García, Jin Wang, and Karl Freed for helpful discussions. This work was supported by the NSF Career Award 013504.

[1] C.L. Brooks III, Curr. Opin. Struct. Biol. **8**, 222 (1998).
[2] J.-E. Shea and C.L. Brooks III, Annu. Rev. Phys. Chem. **52**, 499 (2001).
[3] U. Hansmann and Y. Okamoto, Curr. Opin. Struct. Biol. **9**, 177 (1999).
[4] A.M. Ferrenberg and R.H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).
[5] J.-E. Shea, Y.D. Nochomovitz, Z. Guo, and C.L. Brooks III, J. Chem. Phys. **109**, 2895 (1998).
[6] V. Daggett, Curr. Opin. Struct. Biol. **10**, 160 (2000).
[7] Y. Duan and P.A. Kollman, Science **282**, 740 (1998).
[8] B. Roux, *Implicit Solvent Models* (Marcel Dekker, New York 2001).
[9] J. Guenot and P.A. Kollman, Protein Sci. **1**, 1185 (1992).
[10] J.N. Onuchic, Z. Luthey-Schulten, and P.G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).
[11] G. Hummer, A.E. García, and S. Garde, Phys. Rev. Lett. **85**, 2637 (2000).
[12] J.D. Bryngelson and P.G. Wolynes, J. Phys. Chem. **93**, 6902

(1989).
[13] A. Baumketner and Y. Hiwatari, Phys. Rev. E **66**, 011905 (2002).
[14] M. yi Shen and K. Freed, Biophys. J. **82**, 1791 (2002).
[15] P. Ferrara and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **97**, 10 780 (2000).
[16] D. Poland and H.A. Scheraga, *Theory of Helix-Coil Transitions in Biopolymers; Statistical Mechanical Theory of Order-Disorder Transitions in Biological Macromolecules* (Academic Press, New York, 1970).
[17] *Protein Folding: in Vivo and in Vitro*, edited by J.L. Cleland (ACS Press, Washington, DC, 1993).
[18] Z. Shi, C.A. Olsen, G.D. Rose, R.L. Baldwin, and N.R. Kallenbach, Proc. Natl. Acad. Sci. U.S.A. **99**, 9190 (2002).
[19] L. Wang, T. O'Connell, A. Tropsha, and J. Hermans, Proc. Natl. Acad. Sci. U.S.A. **92**, 10 924 (1995).
[20] S. Huo and J.E. Straub, Proteins: Struct., Funct., Genet. **36**, 249 (1999).
[21] C.T. Klein, B. Mayer, G. Köhler, and P. Wolschann, J. Mol. Struct.: THEOCHEM **370**, 33 (1996).
[22] Y. Levy, J. Jortner, and O. Becker, Proc. Natl. Acad. Sci. U.S.A. **98**, 2188 (2001).
[23] J. Michel, K. Bathany, J.-M. Schmitter, J.-P. Monti, and S. Moreau, Tetrahedron **58**, 7975 (2002).
[24] Ashish and R. Kishore, Bioorg Med. Chem. **10**, 4083 (2002).
[25] A.E. Garcia and K.Y. Sanbonmatsu, Proc. Natl. Acad. Sci. U.S.A. **99**, 2782 (2003).
[26] B.R. Brooks *et al.*, J. Comput. Chem. **4**, 187 (1983).
[27] S. Nosé, Prog. Theor. Phys. **103**, 1 (1991).
[28] M.P. Allen and D.J. Tildesley, *Computer Simulation of Liquids* (Clarendon Press, Oxford, 1986).
[29] C.-L. Lee, G. Stell, and J. Wang, J. Chem. Phys. **118**, 959 (2003).
[30] C.-L. Lee, C.-T. Lin, G. Stell, and J. Wang, Phys. Rev. E **67**, 041905 (2003).
[31] J. Wang, J. Chem. Phys. **118**, 952 (2003).
[32] F. Bardou, J.-P. Bouchaud, A. Aspect, and C. Cohen-Tannoudji, *Lévy Statistics and Laser Cooling: How Rare Events Bring Atoms to Rest* (Cambridge University Press, Cambridge, England, 2002).
[33] S.S. Plotkin and J.N. Onuchic, Q. Rev. Biophys. **35**, 111 (2002).
[34] S.S. Plotkin and J.N. Onuchic, Q. Rev. Biophys. **35**, 205 (2002).
[35] N.D. Socci, J.N. Onuchic, and P.G. Wolynes, J. Chem. Phys. **104**, 5860 (1996).
[36] J.D. Bryngelson and P.G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).
[37] D. Shortle, FASEB J. **10**, 27 (1996).
[38] T.R. Sosnick and J. Trewhella, Biochemistry **31**, 8329 (1992).
[39] W.-Y. Choy *et al.*, J. Mol. Biol. **316**, 101 (2002).
[40] M. Cieplak, T.X. Hoang, and M.S. Li, Phys. Rev. Lett. **83**, 1684 (1999).
[41] M.J. Todd, G.H. Lorimer, and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **93**, 4030 (1996).
[42] M. Shtilerman, G.H. Lorimer, and S.W. Englander, Science **284**, 822 (1999).
[43] Y.-J. Tan, M. Oliveberg, and A.R. Fersht, J. Mol. Biol. **264**, 377 (1996).
[44] C.-Y. Huang *et al.*, Proc. Natl. Acad. Sci. U.S.A. **99**, 2788 (2002).
[45] N.D. Socci, J.N. Onuchic, and P.G. Wolynes, Proteins: Struct., Funct., Genet. **32**, 136 (1998).
[46] H. Nymeyer, A.E. García, and J.N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **95**, 5921 (1998).
[47] S.S. Plotkin and P.G. Wolynes, Phys. Rev. Lett. **80**, 5015 (1998).