Folding dynamics with nonadditive forces: A simulation study of a designed helical protein and a random heteropolymer

Shoji Takada
Department of Chemistry, Kobe University, Kobe, 657-8501, Japan and School of Chemical Sciences, University of Illinois, Urbana, Illinois 61801

Zaida Luthey-Schulten and Peter G. Wolynes
School of Chemical Sciences, University of Illinois, Urbana, Illinois 61801

(Received 8 February 1999; accepted 24 March 1999)

We propose a reduced model of proteins and simulate folding of a designed three helix bundle protein with 54 residues, the dynamics of a random heteropolymer, and the helix formation of a short peptide, up to $\sim 1\ \mu$s, near the estimated lower bound of folding time. The model has explicit backbone atoms, while solvent effects are taken into account via effective potentials. Interactions include two multibody terms; (1) the hydrogen bond strength reflecting the local dielectric constant that is dependent on protein configuration and (2) the hydrophobic force which depends on the local density of peptide atoms imitating the solvent accessible surface area model of hydrophobic force. With this model, all trajectories of a designed protein reach a native like conformation within 0.5 μs although they exhibit some remaining residual fluctuations. On the other hand, a random polymer collapses to several nonspecific compact forms and continues to change its global shape. A 16 residue segment forming a helix in the designed protein does not stably form a helix when it is cleaved, illustrating the effect of nonadditivity. *© 1999 American Institute of Physics.* [S0021-9606(99)51323-8]

I. INTRODUCTION

More than 60 years after the question surfaced, how proteins acquire their unique three-dimensional structures, remains an active area of theoretical and experimental research. Both the diversity of phenomena involved and deep conceptual issues make this problem very difficult. Recently, stimulated to some extent by the energy landscape theory,[1–3] the focus of attention has been shifted to the study of fast folding small proteins.[4–9] Through these experiments the time scales of some elementary folding processes have been uncovered. Helix formation takes place in a few hundred nanoseconds,[6] while formation of a small beta hairpin takes a microsecond.[8] The fastest known time scale for complete folding of a small protein at this moment is of order of tens of μs.[7] Experiments on loop formation together with a simple theoretical model give an estimated lower bound for the folding time of about 1 μs.[10] Recently, Duan and Kollman first performed a full atom simulation of a 36 residue protein subdomain up to 1 μs,[11] in which molten globulelike configurations were reached. But, they could obtain only two trajectories of 1 μs duration using a powerful parallel computer for 4 months. Currently fully atomistic simulations are too slow to address complete folding process with good statistics. Especially, for an all atom simulation it is crucial to include a number of solvent water molecules and the degrees of freedom for the water molecules are sometimes more numerous than those of the protein chain. On the other hand, minimal lattice simulations used to efficiently explore conformation space are often dynamically unrealistic; local features of polypeptide dynamics are inherently missing and both reptation of segments (important in the dense, compact phase) and diffusion of preformed subunits (important at low density and also late in folding) are not usually treated correctly. More elaborate move sets and lattices help in this regard.[12]

Off-lattice, but still simplified models using a reduced description of the chain are more promising in that their dynamics is locally more realistic than the usual lattice models and they can easily be performed up to at least the microsecond and possibly the millisecond regime. Making such models completely realistic especially vis-à-vis relative energies of different configurations presents its own challenges. If simple universal pair potentials are used, the design of easily folded sequences has proved more difficult off lattice than for the lattice case. While sequences that fold to a family of structures with some similarity at low resolution[13] can be found, there is usually a great deal of residual degeneracy[13] like that seen in recent all atom calculation.[11] It is desirable to have a model that can discriminate the native structure with much less fluctuation from the many decoy structures that differ on a large scale.

What features might be necessary to do this? One possibility is the introduction of anisotropic interactions. When done correctly this is tantamount to enriching the number of degrees of freedom in the model. Other attempts to describe such interactions often unwittingly introduce an unphysical anisotropy of space.[14] In this paper we focus on an alternative feature of real proteins in solution that must play a role in their stability and dynamics; the many-body character of solvent and side-chain averaged interactions. These many-body interactions represent effects such as the preorganization of the extra sidechain degrees of freedom of one pair of residues which thereby changes their interaction with a third one. Such many body effects are doubtless present and also can contribute greatly to the thermodynamics and kinetic barriers for folding. A variety of implicit solvent models have been introduced before, some still remaining pairwise additive,[15,16] and others are explicitly many body but are difficult to compute, making long time simulation troublesome.[17]

Here we propose an intermediate level description of proteins, between the usual minimalist models level[13,18–22] and the full atom level, which captures the many-body effects quite efficiently. The basic standpoint used to construct the models is as follows: The model has to be simple enough to allow numerous simulations for at least 1 μs and possibly ~ms with computers currently available. No matter how simple it is, we still require the functional form of the effective potential to be consistent with current physicochemical knowledge. No specific additional interactions that bias the search to the native structure are introduced in the model. Finally, in order for a simple designed protein to fold to a nativelike structure stably, parameters involved in the potentials are varied empirically within the physically reasonable range of the parameters such that they did not deviate too much from experimentally anticipated values.

Our model has a fairly realistic backbone structure, which makes it possible to have a realistic local dynamics, i.e., Ramachandran map and have a sharply varying and anisotropic hydrogen bond between backbone atoms. Solvent molecules are not explicitly treated, but their effects are carefully incorporated into the hydrogen bond model as well as in the hydrophobic force. Each side chain is simply represented as a sphere greatly reducing the number of explicit degrees of freedom. The folding simulation employs Langevin dynamics which allows the model to address the time scale issue. Apparently, the choice of parameters is far from unique. We however note that since successful folding trajectories under physiological temperature were realized for only a limited range of parameters, the parameter set that gives successful folding is not at all a freely-chosen set, but rather passed a severe check.

Our model emphasizes the necessity of including multibody potentials in a reduced representation that does not include coordinates of the water molecules or for side-chain orientations. Even when we assume that fully atomistic models of the protein with explicit water molecules can be approximated by pairwise additive potentials, the multibody nature of the effective interaction arises through reduction of the model to the one that does not include water coordinates explicitly. As stressed by Honig,[23] the free energy cost of breaking a backbone hydrogen bond (HB) strongly depends on the local environment. This cost is small in water, while large in organic solvents. When water molecules are taken into account explicitly, this difference, in principle, comes out through the compensation of hydrogen bonds between two backbone atoms by hydrogen bonds between a backbone atom and a water molecule. On the other hand, in a reduced model that does not include the water molecular coordinates, an *effective* hydrogen bond should be weak at surface of the protein and strong at the hydrophobic core. In other words, the strength depends on the local dielectric constant. An analogous situation applies to the hydrophobic (HP) force, too. The hydrophobic force is a mesoscopic force induced by collective motion of water molecules and can be reproduced, in principle, from pairwise potentials between water molecules and between the solute atoms and water. In a reduced model, however, this should be modeled in terms of highly nonadditive terms such as solvent accessible surface area.[17] Side-chain fixation also will lead to a many body force of a more complex sequence dependent nature.

In this paper, we propose a particular form of multibody potentials in a reduced representation based on the chemical knowledge of the interactions mentioned above, primarily focusing on solvation effects. Without these nonadditive forces, we found the reduced pairwise potential could not simulate folding to the designed native state.

This paper is organized in the following way: Sec. II describes the model in some detail, starting with the chain description and a brief explanation of the Langevin dynamics employed here. Then quite a detailed description of the interaction potential is given. Simulation results and their analysis are presented in Sec. III. First, we show dynamics of a 16 residue peptide, of which sequence corresponds to a helical segment in the three helix bundle protein studied next. We find that at room temperature the helix is sometimes formed, but is easily broken after a while. Next, we present simulations of both a designed three helix bundle protein and a random heteropolymer. The latter has the same amino acid composition as the former, but the sequence is randomly rearranged. Some representative snapshots of folding trajectories are discussed followed by simple time series analysis of many trajectories leading to characteristic time constants. Then, thermodynamic analysis is performed to provide free energy surfaces both in one and two order parameters. At T_F the nativelike state has the right topology (up to two mirror images of the helix alignment), but still shows some large fluctuation particularly at terminal amino acids at folding temperature. Upon further cooling, the designed protein freezes into a nativelike structure, while the random polymer sequence still remains structurally widely distributed leading to a glassy phase. Final structures obtained by many simulated annealing simulations have high correlation (i.e., convergence to the native conformation) for the designed sequence, while their similarity is as low as that between random compact configurations in the case of random polymer. Discussions and conclusion are given in Secs. IV and V, respectively.

II. MODELS

In this section, we describe the model in detail. As is usual for folding simulations, our model description consists of three parts; chain representation, method of time propagation (moving the chain), and the inter-residue potential functions.

A. Chain representation

Our chain includes a very explicit backbone structure with simplified residues (see Fig. 1). The backbone consists of three united atoms per amino acid, NH, $C_\alpha H$, and $C'O$,

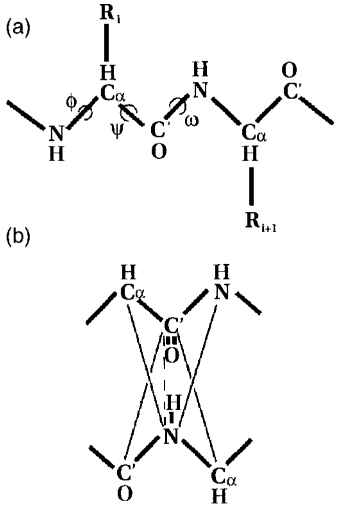

FIG. 1. (a) A schematic figure of chain representation. Coordinates explicitly used in the simulations are drawn in black, others drawn in gray. (b) A schematic figure of hydrogen bond interaction proposed here. Gray dashed line and gray solid lines indicate attractively and repulsively interacting pairs, respectively.

and each side chain residue, except for glycine, is modeled by a bead. Positions of united atoms indicate those of N, C_α, C', and C_β [drawn in black in Fig. 1(a)] and the length of each chemical bond connecting them is constrained, via the SHAKE algorithm, to the value at equilibrium.[24]

B. Langevin dynamics

Since folding is a much slower process than any microscopic motion of chain, it is probably not entirely inappropriate to use the overdamped Langevin equation as a way of moving chain,

$$\zeta_i \frac{d\mathbf{r}_i}{dt} = \mathbf{F}_i + \mathbf{R}_i , \tag{1}$$

where $\mathbf{r}_i$ represents the position of united atoms and $\mathbf{F}_i$ and $\mathbf{R}_i$ are systematic and random forces on the ith united atom, respectively. ζ_i is mass of united atoms times the friction coefficient and is determined by the Stokes Law,

$$\zeta_i = 6\pi a_i^{\text{Stokes}} \eta, \tag{2}$$

where a_i^{Stokes} is the effective radius of each united atom and is chosen as the van der Waals (vdW) radius plus radius of the water molecule, 1.4 Å. η is the viscosity of water. The random force $\mathbf{R}_i(t)$ used here is Gaussian and white with mean zero and variance,

$$\langle \mathbf{R}_i(t)\mathbf{R}_j(t') \rangle = 2\zeta_i k_B T \delta_{ij} \delta(t-t') \mathbf{1} \tag{3}$$

that comes from the fluctuation-dissipation theorem. Here the brackets $\langle\,\rangle$ denote an ensemble average and $\mathbf{1}$ is a 3×3 unit matrix.

In the actual simulation, a move for each finite time step h includes two steps. First, an unconstrained move due to the force $\mathbf{F}+\mathbf{R}$, and then the SHAKE procedure. The former is, as usual,

$$\mathbf{r}_i^{\text{unc}}(t+h) = \mathbf{r}_i(t) + \mathbf{F}_i(t)h/\zeta_i + \Delta\mathbf{R}_i(h), \tag{4}$$

where $\Delta\mathbf{R}_i$ is a random kick for the time interval h induced by $\mathbf{R}_i(t)$, of which the average is zero and the variance derived from Eq. (3) is

$$\langle \Delta\mathbf{R}_i(h)\Delta\mathbf{R}_j(h) \rangle = 2\frac{k_B T}{\zeta_i} h \delta_{ij} . \tag{5}$$

To satisfy the constraint for bond lengths, we use the standard SHAKE algorithm. It consists of a move of each pair of bonding atoms so that this pair satisfies the constraint and repeats this for every pairs until all constraints are simultaneously satisfied within some tolerance.

C. Potentials

Systematic forces $\mathbf{F}_i$ are, as usual, the derivatives of the potential energy function, i.e., $-\partial V/\partial \mathbf{r}_i$. Potentials are composed of many terms that are classified into two categories, one of local V_{lc} terms and the other of nonlocal terms V_{lc} along the chain

$$V = V_{\text{lc}} + V_{\text{nonlc}} . \tag{6}$$

The local potential V_{lc} consists of many terms,

$$V_{\text{lc}} = V_\theta + V_\omega + V_\phi + V_\psi + V_{\text{vdW-lc}} + V_{\text{chiral}} \tag{7}$$

of which most have quite standard functional forms. The bond angle potential V_θ is harmonic with respect to the bond angle θ_i constraining the angle to its equilibrium value θ_{0i},

$$V_\theta = \sum \frac{1}{2} k_{\theta_i} (\theta_i - \theta_{0i})^2, \tag{8}$$

where k_{θ_i} is the harmonic constant and the sum is over all bond–bond angles. V_ω keeps the peptide bond close to planar,

$$V_\omega = \sum \frac{1}{2} \epsilon_\omega (1 + \cos \omega_i), \tag{9}$$

where ω_i is a dihedral angle that reflects rotation around the ith peptide bond and ϵ_ω is an energetic cost for having *cis*-conformation relative to the *trans*-conformation. V_ϕ and V_ψ are potentials for conformation change with respect to ϕ and ψ, respectively, and have standard forms with the threefold symmetry,

$$V_\phi = \sum \frac{1}{2} \epsilon_\phi (1 + \cos 3\phi_i) \tag{10}$$

and

$$V_\psi = \sum \frac{1}{2} \epsilon_\psi (1 + \cos 3\psi_i), \tag{11}$$

where energy parameters ϵ_ϕ and ϵ_ψ determine barriers for conformational change and thus are responsible for the chain rigidity. $V_{\text{vdW-lc}}$ is a vdW potential only for so-called 1–4 pairs, namely, pairs that are connected through three covalent bonds,

$$V_{\text{vdW-lc}} = \sum_{1-4\ \text{pair}} \epsilon_{\text{vdW-lc}} \left[\left(\frac{a_{\text{lc},ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{a_{\text{lc},ij}}{r_{ij}} \right)^{6} \right], \tag{12}$$

where $a_{\text{lc},ij} = a_{\text{lc},i} + a_{\text{lc},j}$. As is in the usual full atomic model, the vdW interaction between 1–4 pair is set weaker

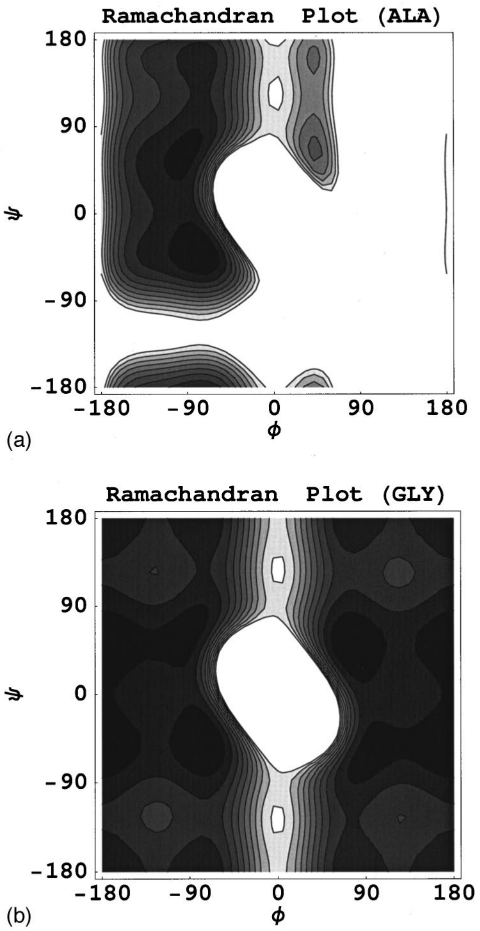

FIG. 2. The Ramachandran potential by the current model. The energy interval for the contours is 0.4 kcal mol^{-1}. Dark region has lower energy. (a) For nonglycine amino acids and (b) for glycine.

than other vdW interactions. V_{chiral} is a simple energy cost function that has large positive energy ϵ_{chiral} when the chirality of any C_α is violated. Whether the chirality is violated or not can be easily detected by the sign of the triple product $P_3=(\mathbf{r}_{NC_\alpha}\times\mathbf{r}_{C_\beta C_\alpha})\cdot\mathbf{r}_{C'C_\alpha}$, where $\mathbf{r}_{ij}\equiv\mathbf{r}_i-\mathbf{r}_j$. Therefore,

$$V_{\text{chiral}}=\sum_I v_{\text{chiral},I}, \tag{13}$$

where I stands for α carbon and

$$v_{\text{chiral},I}=\begin{cases}0 & \text{when } P_3<0\\ \epsilon_{\text{chiral}}P_3 & \text{when } P_3>0\end{cases}. \tag{14}$$

Many parameters used in the local part of the potentials are determined so that the local stability and flexibility of the peptide are close to those of real protein *in solution*. Using these terms, we plot the derived Ramachandran potential for ϕ and ψ angles. Namely, we compute the energy of tripeptide GLY-X-GLY changing ϕ and ψ angles of X, where X is the target amino acid. We then compared it with the known Ramachandran potential.[25] We depict our Ramachandran potentials with the parameters actually used for nonglycine residues and glycine in Fig. 2. For nonglycine residues, a low energy region exists near the helical conformation as well as the extended form. Glycine has a much wider accessible area as is expected.

Next, we describe the nonlocal part of the potentials. This consists of three terms,

$$V_{\text{nonlc}}=V_{\text{vdW}}+V_{\text{HB}}+V_{\text{HP}}. \tag{15}$$

V_{vdW} has exactly the same form as that between 1–4 pairs, $V_{\text{vdW-lc}}$, but with different parameter values, ϵ_{vdW} and $a_{ij}=a_i+a_j$. We note that van der Waals radii used in local and nonlocal interactions have different meanings. For 1–4 pairs, the major van der Waals interaction arises from overlap between two atoms (instead of united atoms) that are actually connected via 1–4 distance, i.e., N, C_α, C', and C_β. Therefore, for 1–4 interaction the radii of these atoms are more appropriate than the effective radii of united atoms. On the other hand, nonlocal interactions mainly occur on the opposite side from the backbone, and thus we use an effective van der Waals radius for an united atom in the nonlocal part. Actually, these different values of vdW radii were necessary to yield reasonably good Ramachandran maps and stable helix simultaneously. Local and nonlocal van der Waals potentials primarily affect the former and the latter, respectively.

The hydrogen bond interaction is one of the most important parts of our model. Three important ingredients are involved. First, our hydrogen bond model is appropriately anisotropic. In standard full atom models, the angle dependence of the hydrogen bond is realized by introducing the potential as a function of bond distance and angle,[26] which is somewhat expensive from the point of view of computation. Instead of this, we introduce a particular combination of attractive and repulsive forces that result in an anisotropic hydrogen bond interaction. More concretely, an attractive interaction is introduced between $C'O$ and NH, while a repulsive force of half as strong as the attractive one is introduced between NH and NH, between NH and $C_\alpha H$ between $C_\alpha H$ and $C'O$, and between $C'O$ and $C'O$ [see Fig. 1(b)]. A second feature we introduce is a context dependence of the hydrogen bond interaction as mentioned above. Honig and his group estimate that the free energy changes in breaking hydrogen bonds in water and in organic solvent are 0.5 and 3.8 (in kcal/mol), respectively.[23] This difference is nonnegligible although these quantities may include significant errors because this estimate is quite indirect. In globular proteins the local environment for an atom is well monitored by local dielectric constant or the local density of peptide atoms; an organic environment corresponds to low dielectric constant and high local density of peptide atoms. In our model, the hydrogen bond strength depends on the local density of peptide atoms. We notice that exposed NH and CO can make hydrogen bonds with water molecules as well. Thus, both the exposed [NH OC′] and [NH···OC′] basically make hydrogen bonds with either a peptide atom or a water molecule. In this respect, only the buried [NH OC′] cannot make any H bond. Thus, the third feature we introduce is that the hydrogen bond potential should have positive energy for a buried non-H bonded [NH OC′] pair. For this purpose, we add a penalty term that gives a positive energy for a high density of peptide atoms. In summary our hydrogen bond interaction model can be written as

$$V_{\mathrm{HB}}=\epsilon_{\mathrm{HB}}\sum_{ij(\mathrm{s.t.}I\geqslant J+3)} S_{\mathrm{HB},IJ}u_{\mathrm{HB}}^{(a,r)}(r_{ij}) + \epsilon_{\mathrm{HB}}c_{\mathrm{HB},c}\sum_{I} S_{\mathrm{HB},c,I}, \tag{16}$$

where I and J represent amino acids in which ith and jth atoms are belonging to, respectively. ϵ_{HB} is the (constant) hydrogen bond strength, $S_{\mathrm{HB},IJ}$ represents local density dependence of actual hydrogen bonds, and $u_{\mathrm{HB}}^{(a,r)}(r_{ij})$ are the distance dependent part. The superscripts (a) and (r) mean attractive and repulsive, respectively. Depending on the atom type of i and j, either an attractive or repulsive form is chosen [see Fig. 1(b)],

$$u_{\mathrm{HB}}^{(a)}(r_{ij})=60\left[\frac{1}{12}\left(\frac{\sigma_{\mathrm{HB}}}{r_{ij}-r_{\mathrm{HB},0}}\right)^{12}-\frac{1}{10}\left(\frac{\sigma_{\mathrm{HB}}}{r_{ij}-r_{\mathrm{HB},0}}\right)^{10}\right] \tag{17}$$

and

$$u_{\mathrm{HB}}^{(r)}(r_{ij})=30\left[\frac{1}{10}\left(\frac{\sigma_{\mathrm{HB}}}{r_{ij}-r_{\mathrm{HB},0}}\right)^{10}\right], \tag{18}$$

where r_{ij} is the distance between the ith and jth atom, σ_{HB} and $r_{\mathrm{HB},0}$ are parameters that mean the length scale for hydrogen bond and distance shift. The latter offset is needed because the hydrogen bond is made between H and O, while in our model united atoms are located on the sites of N and C. The local density dependence $S_{\mathrm{HB},IJ}$ is a linear combination of two terms, $S_{\mathrm{HB},IJ}=(S_{\mathrm{HB},I}+S_{\mathrm{HB},J})/2$, each of which is defined by

$$S_{\mathrm{HB},I}=S_{\mathrm{HB}}\left(\sum_{K} u_{\mathrm{HP}}(r_{IK}),\eta_{\mathrm{HP,min}},\eta_{\mathrm{HB,max}}\right), \tag{19}$$

where $\Sigma_K u_{\mathrm{HP}}(r_{IK})$ monitors the local density on the Ith amino acid and r_{IK} is the distance between α carbons of the Ith and Kth residues. The function form of $u_{\mathrm{HP}}(r)$ will be specified below. The function S_{HB} is a smooth switching function and has an explicit form of

$$S_{\mathrm{HB}}(x,x_{\mathrm{min}},x_{\mathrm{max}}) = \begin{cases} 1 & \text{when } x>x_{\mathrm{max}} \\ \left(1+\cos\pi\dfrac{x_{\mathrm{max}}-x}{x_{\mathrm{max}}-x_{\mathrm{min}}}\right)\Big/2 & \text{when } x_{\mathrm{min}}\leqslant x\leqslant x_{\mathrm{max}} \\ 0 & \text{when } x<x_{\mathrm{min}}. \end{cases} \tag{20}$$

The second term in Eq. (16) is the penalty term where $c_{\mathrm{HB},c}$ represents the ratio of the penalty to the first term and $S_{\mathrm{HB},c,I}$ represents local density. $S_{\mathrm{HB},c,I}$ is defined by the same equation as Eq. (19), but with different parameters $\eta_{\mathrm{HB},c,\mathrm{min}}$ and $\eta_{\mathrm{HB},c,\mathrm{max}}$.

Finally, we describe the hydrophobic interaction. This interaction is sometimes modeled in terms of solvent accessible surface area (ASA) which is decidedly nonadditive.[17] It is, however, somewhat too computationally time-consuming for our present purpose of long time scale simulation. More seriously, the exact analytical calculation of ASA suffers from singularity problems at some configurations. In applications for molecular dynamics, this unstable behavior represents a severe deficiency. We propose a model somewhat similar to the ASA model, but which is written in terms of the local density of peptide atoms. We only consider the HP interaction among C_α atoms and among C_β atoms. The latter depends on the type of amino acid, while the former is independent of the type of amino acid. The explicit form of our HP model is

$$V_{\mathrm{HP}}=-\epsilon_{\mathrm{HP}}^{(\alpha)}\sum_{I} S_{\mathrm{HP}}\left(\sum_{K} u_{\mathrm{HP}}(r_{IK}),\eta_{\mathrm{HP}\alpha}\right) -\sum_{\mu} \epsilon_{\mathrm{HP},\mu}^{(\beta)}S_{\mathrm{HP}}\left(\sum_{\nu} u_{\mathrm{HP}}(r_{\mu\nu}),\eta_{\mathrm{HP}\beta}\right), \tag{21}$$

where I and K represent α carbons and μ and ν represent C_β atoms, $\epsilon_{\mathrm{HP}}^{(\alpha)}$ and $\epsilon_{\mathrm{HP},\mu}^{(\beta)}$ are the hydrophobic energy parameters for the C_α and side chains, and the function u_{HP} represents the distance dependence of the interaction that is defined as a smooth switching function as follows:

$$u_{\mathrm{HP}}=\begin{cases} 1 & r<\sigma_{\mathrm{HP1}} \\ 1/2\left(1+\cos\pi\dfrac{r-\sigma_{\mathrm{HP1}}}{\sigma_{\mathrm{HP2}}-\sigma_{\mathrm{HP1}}}\right) & \sigma_{\mathrm{HP1}}<r<\sigma_{\mathrm{HP2}} \\ 0 & r>\sigma_{\mathrm{HP2}}. \end{cases} \tag{22}$$

In the expression of the HP interaction, Eq. (21), $\Sigma_K u_{\mathrm{HP}}(r_{IK})$ monitors the local density at the site I. The function S_{HP} represents the buriedness of the atom defined as

$$S_{\mathrm{HP}}(x,x_{\mathrm{max}})=\begin{cases} 1 & \text{when } x>x_{\mathrm{max}} \\ \cos\dfrac{\pi}{2}\dfrac{x_{\mathrm{max}}-x}{x_{\mathrm{max}}} & \text{when } x\leqslant x_{\mathrm{max}}. \end{cases} \tag{23}$$

This term is unity when the atom is sufficiently buried, and close to zero for an atom well exposed to the solvent. Since this term is expressed in terms of the local density, we compute it first and store it for later use. Computational time for this force is nearly as short as for typical two-body forces. Actually, the expression for the force is quite similar to that of a simple two-body contact model but in front of an ordinary two-body force expression, there is an additional prefactor that depends on density. This factor weakens the attractive force when the local density is high. That is, the three-body effect has the opposite sign of the two-body energy. We also note that, compared with ordinary two-body contact energies, we only need as many energy parameters as types of amino acids (20 plus a few, instead of its square). More importantly, these quantities are experimentally measurable. The values of all the parameters included in the model are listed in Tables I and II.

D. Studied model peptides

We performed MD simulations for three systems that are mutually related to each other. In all cases, only three types of amino acid are used, namely, ALA(=A), SER(=S), and GLY(=G), as representatives of hydrophobic residues, polar residues, and turn-making residues, respectively. For the de-

TABLE I. Parameters I. $a^{\mathrm{Stokes}} = a_{\mathrm{non\text{-}lc}} + 1.4$ Å.

Bond lengths	r (Å)	
C_α–C'	1.52	
C_α–N	1.45	
C'–N	1.33	
C_α–C_β	1.8	
Bond angles	θ_{0i} (deg)	k_{θ_i} (kcal mol^{-1}/rad^2)
$\angle$(N–C_α–C')	111.6	200
$\angle$(C_α–C'–N)	117.5	200
$\angle$(C'–N–C_α)	120.0	200
$\angle$(N–C_α–C_β)	110.0	200
$\angle$(C'–C_α–C_β)	110.0	200
vdW radii	a_{lc} (Å)	$a_{\mathrm{non\text{-}lc}}$ (Å)
C_α	1.48	1.85
C'	1.70	2.00
N	1.32	1.65
C_β	2.52	2.52

signed sequence, we took the same strategy as Guo and Thirumalai.[27] The *de novo* designed sequence of DeGrado[28] is further simplified to use only three types of amino acids. Moreover, the three helix bundle is obtained by cutting one helix and turn from DeGrado's four helix bundle. Namely, each helix is modeled as

$$\mathrm{SSASSAASSASSAASS} \tag{24}$$

while a turn is designed as G–G–G. Then we connected three helices and two turns to make a three helix bundle protein. The total length of this protein is 54 residues. For comparison we randomly mutated this 54 residue proteins getting a random heteropolymer with the same amino acid composition as the designed protein. In particular, we use the sequence

$$\begin{aligned}&\mathrm{SSAGSGSSASSAASSSGSAASASASSAG}\\&-\mathrm{SASASSASSSSSSSASGASSAASAAG}.\end{aligned} \tag{25}$$

This has sequence homology of 38.8% to the designed one, which is as low as that expected, 43.2%, for ideally random shuffling. Finally, helix formation and deformation is investigated for an isolated 16 residue peptide defined in (24).

TABLE II. Parameters II.

	Local potentials	
ϵ_ω	40 kcal mol^{-1}	
ϵ_ϕ	0.45 kcal mol^{-1}	
ϵ_ψ	0.45 kcal mol^{-1}	
$\epsilon_{\mathrm{vDW\text{-}lc}}$	0.3 kcal mol^{-1}	
$\epsilon_{\mathrm{chiral}}$	10 kcal mol^{-1}	
	Nonlocal potentials	
vdW	ϵ_{vdW}	0.165 kcal mol^{-1}
HB	ϵ_{HB}	2.8 kcal mol^{-1}
	$r_{\mathrm{HB},0}$	1.43 Å
	σ_{HB}	2.0 Å
	$\eta_{\mathrm{HB,min}}$	1.0
	$\eta_{\mathrm{HB,max}}$	9.0
	$c_{\mathrm{HB},c}$	0.5
	$\eta_{\mathrm{HB},c,\mathrm{min}}$	3.0
	$\eta_{\mathrm{HB},c,\mathrm{min}}$	12.0
HP	$\epsilon_{\mathrm{HP}}^{(\alpha)}$	1.12 kcal mol^{-1}
	$\epsilon_{\mathrm{HP,H}}^{(\beta)}$	2.8 kcal mol^{-1}
	$\epsilon_{\mathrm{HP,P}}^{(\beta)}$	0.0 kcal mol^{-1}
	$\eta_{\mathrm{HP}\alpha} = \eta_{\mathrm{HP}\beta}$	12.0
	σ_{HP1}	3.5 Å
	σ_{HP2}	9.5 Å
	MD parameters	
η	$0.01P = 0.1439$ kcal mol^{-1} Å^{-3} ps	viscosity of water
T	250 K, 300 K, 350 K, 400 K, 500 K	temperature
h	0.01 ps	time increment
SHAKE tolerance	0.01	relative to bond length

III. RESULTS

This section starts with describing the parameter optimization scheme and simulation protocol used for the following analysis. Then, we present simulation results of the 16 residue peptide (HLX16) studying the helix–coil transformation. Finally, simulations results of the designed helix bundle protein (PRO54) and the random heteropolymer (RHP54) are studied both from the aspects of kinetics and thermodynamics.

A. Parameters determination and simulation protocol

As mentioned earlier, parameters in the local part of potential are determined so that the resultant Ramachandran plot of the potential function is close to the available results obtained in solution. We show the Ramachandran plot in Figs. 2 both for non-GLY and GLY residues. Parameters in the nonlocal potential are more difficult to choose. We first determined the length scale parameters in the hydrogen bond interaction, $r_{\mathrm{HB},0}$ and σ_{HB}, using the criteria that the helix is stably formed for $(\mathrm{ALA})_{10}$ with the strengthened hydrogen bond interaction and that the length scale itself is acceptable from chemical knowledge. The HP interaction between two methane solutes is known to reach up to about 8 Å (Ref. 29) and thus we decided for the HP length scale σ_{HP2} to use a somewhat larger value than that representing the average size of non-GLY residues. We estimated that a typical value of local density $\Sigma_K u_{\mathrm{HP}}(r_{IK})$ in the core of a three helix bundle structure is about 9. This is used to choose the η's. Energetic parameters are more crucial in obtaining native like structure. We first fixed the hydrophobic energy of polar residue interactions to equal zero. The other major energy parameters, namely $\epsilon_{\mathrm{HB}}, \epsilon_{\mathrm{HP}}$'s are determined largely empirically. Namely, with the constraint of $\epsilon_{\mathrm{HB}} < 4.0$ kcal/mol, $\epsilon_{\mathrm{HP}} < 4.0$ kcal/mol, we repeated folding simulations of the designed sequence from random coil structures at 300 K and checked if trajectories go into a nativelike conformation or not. We end up with a set of parameters listed in Tables I and II.

With this set of parameters, we performed a series of simulations for three systems. For all cases, we started simulations from a random coil structure, which is constructed by randomly choosing (ϕ, ψ) angles and rejecting structures that have overlap of peptide atoms. For the 16 residue peptide

(HLX16), we collected 10 trajectories at 300 K, each of which runs for 0.2 μs. Major helix formation and destruction were found typically once or twice in each run and so as a whole about 20 transformations were sampled. We used a bit longer time sampling for RHP54; 10 trajectories of 0.4 μs runs at 300 K. In each trajectory, we picked up the lowest energy configuration and did a simulated annealing quench from it. For the latter, starting from 300 K, the temperature was lowered every 1.5 ns by 5 K until reaching 0 K. For the RHP54, the sampling is somewhat poorer than we would like although we think it is sufficient for the analysis below. Sampling for the designed helical protein (PRO54) was performed more exhaustively. At 300 K, folding simulations from random coil structures were repeated 20 times, in which 5 trajectories run up to 1 μs and the other 15 trajectories up to 0.4 μs. In order to sample a bit higher energy conformation, we also did the simulations at higher temperatures, 6 trajectories of 0.4 μs long at 350 K, 2 trajectories of 0.4 μs long at 400 K, and 2 trajectories of 0.4 μs long at 500 K. In the same way as in RHP54, we did simulated annealing quenches from the lowest energy structure in the 300 K trajectories to find the native structure. For all the above cases, atomic configurations at every 0.1 ns are stored in individual files and are used for the following analysis.

As a measure of distance to the native structure, we introduce a nativeness order parameter

$$Q=\sum \exp\left[-\left(\frac{r_{IJ}-r_{IJ}^{\mathrm{nat}}}{3}\right)^2\right] \Bigg/ [\text{number of summed pairs}], \tag{26}$$

where, r_{IJ} and r_{IJ}^{nat} are the α-carbon distances (in Å) in a given structure and in the native structure (the lowest energy structure for RHP 54), the summation is taken over all nonlocal pairs, i.e., pairs with $|I-J|\geqslant 3$, and the denominator is the number of such pairs normalizing Q. In the case of PRO54, native topology turned out to fall into the range $0.55\leqslant Q\leqslant 1$, while disordered compact structures typically have $Q\sim 0.4$.

The thermodynamic study below uses standard histogram methods[30] to compute the free energy profile and thermodynamic quantities. We use both the single histogram (SH) method and multiple histogram method (WHAM that stands for the weighted histogram analysis method). In case of the SH method, the histogram is constructed by sampling phase space at 300 K alone. We can compute the free energy profile (i.e., potential of mean force) $F(X)$ with respect to X, where X can be any function of the peptide atom coordinates, such as radius of gyration R_g, nativeness Q, number of helical amino acids $\#\alpha$. The formula can be written as

$$\exp[-\beta_2 F(X)]=\frac{\Sigma_V N(X,V)e^{(\beta_1-\beta_2)V}}{\Sigma_{X,V}N(X,V)e^{(\beta_1-\beta_2)V}}, \tag{27}$$

where $N(X,V)$ are the number of snapshots that have X and V in the range of $X\sim X+\Delta X$ and $V\sim V+\Delta V$ (ΔX and ΔV are the widths of bins). β is, as usual, $1/(k_BT)$, $T_1=300$K, and T_2 is the temperature at which we want the free energy profile. The WHAM is a sophisticated variant of the SH method and can efficiently utilize sampling with different conditions (temperature for now). The precise formula are found in Ref. 30.

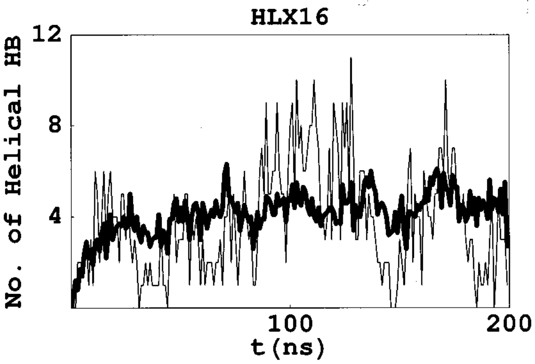

FIG. 3. Number of formed hydrogen bonds as a function of time for HLX16. Thin curve is for a single trajectory and thick curve is an average over trajectories.

B. Helix–coil transformation in a 16 residue peptide (HLX16)

First, we look at the α helix formation as a function of time. The degree of helix formation is monitored by the number of helical hydrogen bonds $\#\alpha$, which is depicted in Fig. 3 as a function of time. Here, α helical hydrogen bonds are defined as those between the Ith and $I+4$th amino acids and we uses Kabsch–Sander's criteria for H bonding.[31] In the figure, the thin curve represents results of a single trajectory, where an incomplete helix is formed from 80 ns to 130 ns, as well as around 175 ns. Sharp growth at 80 ns and ~170 ns implies modestly cooperative formation of helix. The thick curve in the figure is the average over 10 trajectories, which grows within 25 ns at the very beginning. The time scale for the helix formation in this model is about 15 ns, somewhat fast compared with the experimental result. The number of hydrogen bonds formed fluctuates around five, which means a bit less than half the hydrogen bonds are formed on average.

Using of the SH method, we computed the potential of mean force $F(\#\alpha)$ at $T=250$ K,300 K,350 K which are drawn in Fig. 4. First of all, we note that the curve is quite

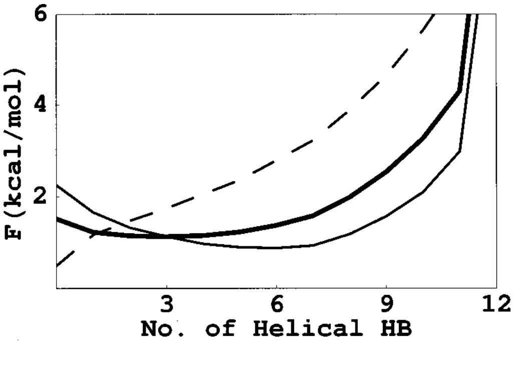

FIG. 4. Free energy curves in terms of number of formed helical hydrogen bonds for HLX16. Thin solid curve, thick solid curve, and dashed curve are results at 250 K, 300 K, and 350 K, respectively.

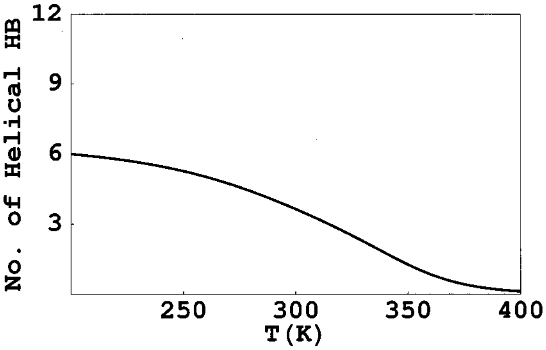

FIG. 5. Average number of formed helical hydrogen bonds as a function of temperature for HLX16.

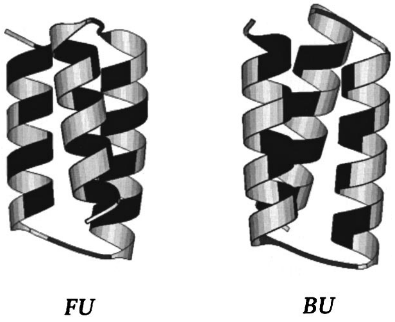

FIG. 6. A pair of our native structures, FU and BU for PRO54. [Drawn with MolScript (Ref. 32).]

flat, implying the stability of helix relative to the coil structure is near neutral. During the trajectory the difference in the total energy is as large as 20 kcal/mol between the instance of mostly helical structure and the instance of mostly random coil structures. This implies that almost perfect energy-entropy compensation is accomplished. (We note that the *energy* in our model includes entropic contribution from solvents in real protein in solution, whereas the *entropy* here means chain entropy.) As expected, the helix becomes more stable at low temperature and less stable at high temperature. The temperature dependence is shown in a different way in Fig. 5, where the average number of helical hydrogen bonds is plotted against temperature. A sigmoidal, but not very sharp, curve is found as expected. We also investigated another sequence $(\mathrm{ALA})_{16}$ (unpublished), which shows a bit more stable helix, but exhibits qualitatively the same results.

Here, we note, in advance, that this same segment will make a rather stable helix when it appears in the 54 residue long peptide. This illustrates that in our model, as in the laboratory, secondary structure elements of proteins, when cleaved, do not necessarily form the same secondary structure as stably as in the intact system. In particular, the backbone amide H and carbonyl O need to form hydrogen bonds in the hydrophobic core in order that the polar interactions in the backbone are canceled out. This is what makes the helix stable in the intact protein. On the other hand, hydrogen bonds are not very crucial for a surface exposed group. As Honig stressed, this is an important feature of real proteins and we believe that use of a context dependent hydrogen bond model is indeed necessary for realizing this feature if explicit water molecules are not included in the model.

C. The designed helical protein (PRO54) and the random heteropolymer (RHP54): Kinetics

Now we present the results of the designed helical protein (PRO54) and the random heteropolymer (RHP54), comparing them from a kinetic viewpoint. In the simulation of PRO54 at 300 K, all trajectories fold into native like three helix bundle structures. But, we find two possible alignments of three helices described as follows: suppose that we write a character ''U'' on a plane by the N-terminal helix, turn, and the middle helix, and then we see that the third helix (C-terminal helix) may be either in front of or behind the plane. These two are distinct alignments, and we call them ''FU'' (front of U) and ''BU'' (back of U), respectively. Those are depicted in Fig. 6, where dark segments are ALA, gray GLY, and white SER. Out of 20 trajectories, about half (11/20) fold into FU and half (9/20) BU. When we carry out the simulated annealing calculations starting from both structures at 300 K, the lowest energy found in each alignment is almost identical, implying that our model cannot discriminate these two structures. Looking at the pictures of these structures, we notice that both have a maximal number of helical hydrogen bonds and, all the amino acids embedded in the core are the same, i.e., ALA in both cases, and that turns are composed of very flexible amino acids (GLY) alone. With such a reduced representation of amino acid coding, namely 3 letters, it may be inherently difficult, if barely possible, to distinguish these two alignments by a distinct energy difference. This may illustrate the necessity of using more than three types of amino acids for better design of unique structures.[33] For a given structure, the nativeness Q is calculated with reference to both structures (Q_{FU} and Q_{BU}) and the larger value is taken, namely, the definition of Q is modified to be $Q=\mathrm{MAX}(Q_{\mathrm{FU}},Q_{\mathrm{BU}})$. Doing the simulated annealing calculation, we looked for the lowest energy structure in the RHP54, too. As expected, this is very difficult computationally much as for the traveling salesman problem. Our purpose here is not to precisely locate the ground state of the RHP, but to prepare a reference structure which has an energy (not necessarily structure) close to the real ground state. The lowest energy state found is used as the reference ''native'' structure for the analysis below.

Figure 7 illustrates two representative folding courses for PRO54; one is fast direct folding and the other is indirect folding via transient misfolded structures. For the first trajectory [Fig. 7(a)], global nonspecific compaction comes with partial helix formation within a few tens of nanoseconds. Helix formation proceeds further via parallel alignment of these fast forming helices (80 ns). An almost native like structure is already attained by 120 ns, although the angles between helices are not yet right and the terminus of each helix is partially melted. This process is followed by a slow adjustment of angles between helices to reach a native structure (BU shape). For the second illustrative trajectory [Fig. 7(b)], the first and third helices are incompletely formed within 80 ns. After a while, these two segments come close but make *wrong*, i.e., non-native contacts between the two

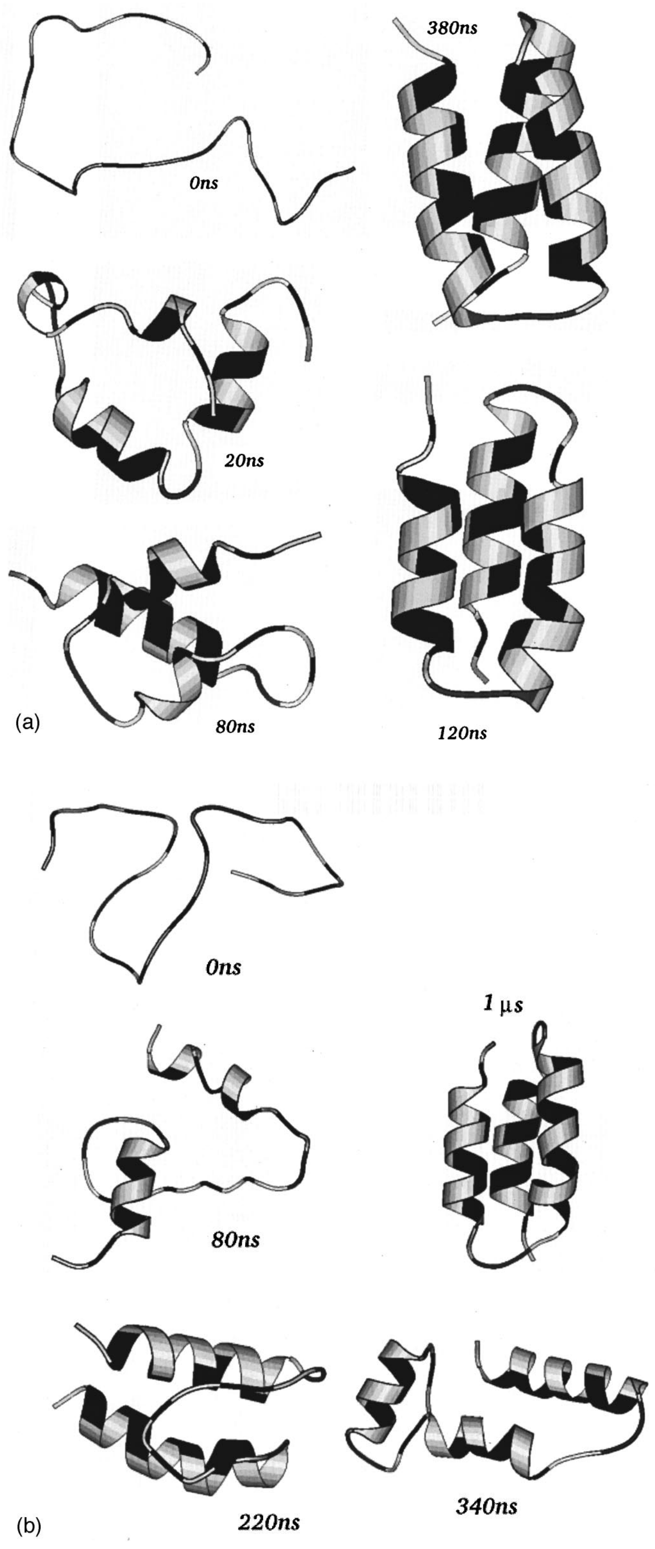

FIG. 7. Representative snapshots of fast folding trajectories for PRO54. [Drawn with MolScript (Ref. 32).] (a) Along the time axis, the nativeness Q is 0.10, 0.52, 0.51, 0.75, 0.87; (b) Along the time axis, the nativeness Q is 0.15, 0.49, 0.52, 0.41, 0.62.

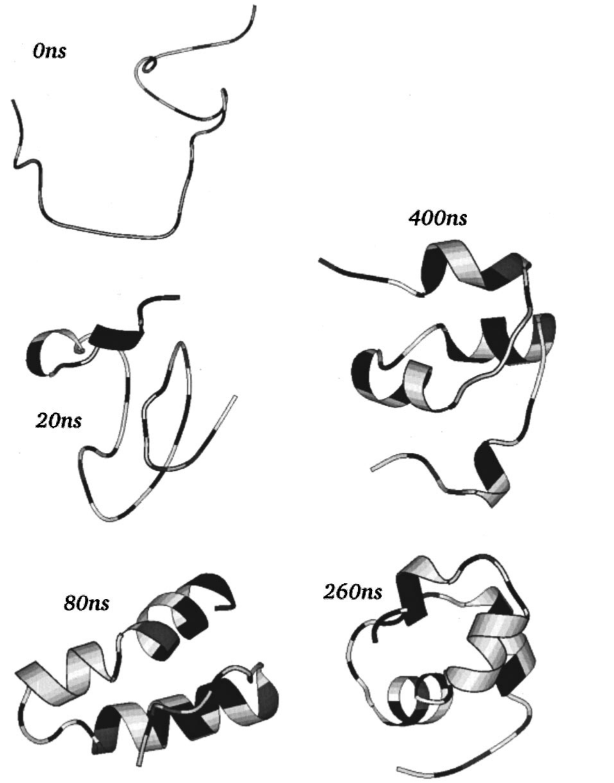

FIG. 8. Representative snapshots of trajectories for RHP54 [drawn with MolScript (Ref. 32)]. Along the time axis, the nativeness Q is 0.13, 0.39, 0.53, 0.43, 0.42.

helices with random coil in between. The first and third helices are antiparallel at 0.22 μs, while they are parallel in the native state. This misfolded structure is quite stable lasting for more than 100 ns. Escape from this misfolded metastable topology occurred through opening this wrong compact form at ~0.34 μs leading to the nativelike structure at 0.5 μs. These two trajectories are qualitatively very different and clearly exemplify a diversity of folding routes in fast folding events.

Trajectories for the RHP54 are quite different at 300 K, which is illustrated in Fig. 8. Initial compaction accompanied by partial helix formation until ~60 ns resembles that found in PRO54. After this stage, however, the peptide appears to find better hydrophobic contacts but at the cost of the hydrogen bonds. The peptide in general keeps transforming its shape without becoming trapped by any deep minimum at 300 K. It is noted that all trajectories look different one from another in RHP54.

The change of radius of gyration $R_g = \sqrt{\Sigma_I r_I^2/N}$ with time is plotted both for PRO54 and RHP54 in Fig. 9 (thin curves) for a couple of trajectories where $N=54$. The peptide is located so that the center of mass is at the origin of the coordinate system. The average over trajectories is plotted as a thick curve. After initial compaction, the averaged R_g is not very different for PRO54 and RHP54. For individual runs, in contrast to the small fluctuations in time and small deviations between different trajectories of PRO54, RHP54 has significantly larger fluctuations. The hydrophobic core is fragile and opening of this core leads to global structural changes in the case of RHP54. For PRO54, a least square fit to this average with a single exponential function is made yielding collapse time constants of ~30 ns.

Next, we show α helix formation as a function of time in Fig. 10 both for PRO54 and RHP54. In the same way as the

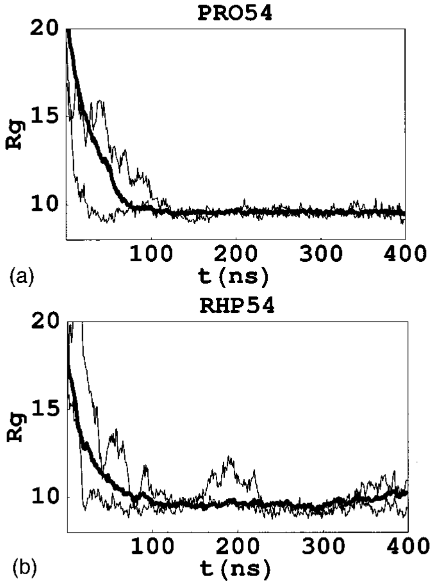

FIG. 9. Radius of gyration as a function of time. Two individual cases are drawn with thin curves, the average over all trajectories is in thick curve. (a) PRO54, (b) RHP54.

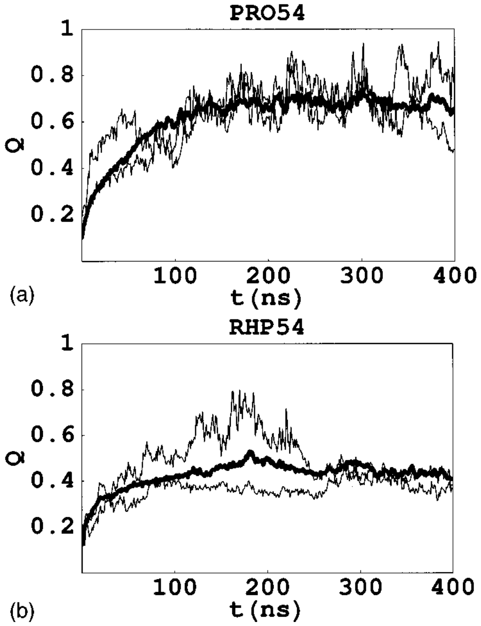

FIG. 11. Global nativeness Q as a function of time. Two individual cases are drawn with thin curves, the average over all trajectories is in thick curve. (a) PRO54, (b) RHP54.

former subsection, this is monitored by measuring the number of hydrogen bonds between the Ith and $I+4$th sites. The criteria for hydrogen bond formation is that of Kabsch and Sander.[31] We find about 26 hydrogen bonds on average in PRO54 and 20 hydrogen bonds in RHP54. It is interesting to note that two thin curves (individual trajectories) for RHP54 exhibit an overshooting of the helical content. From 100 ns to 200 ns, the helical content is somewhat larger than those found later. As illustrated by snapshots in Fig. 8, after making many hydrogen bonds, the peptide tries to stabilize hydrophobic interaction at the cost of helical hydrogen bonds. Note that helical hydrogen bonding is made via local contacts taking less time than hydrophobic contacts, predominantly nonlocal, and therefore slower. A similar phenomenon has been observed in lactoglobulin by Goto's group.[34] No such overshoot phenomenon is observed in PRO54. This illustrates the so-called consistency principle between secondary and tertiary interactions postulated by Gō for natural protein long ago[18] suggesting that this is indeed an important way of realizing minimal frustration.[1] In the same way as above, we plot the average over trajectories as a thick curve. We also fit this average to a single exponential function leading to helix formation time constant ~40 ns. This is several times as fast as that measured from experiments for larger proteins.

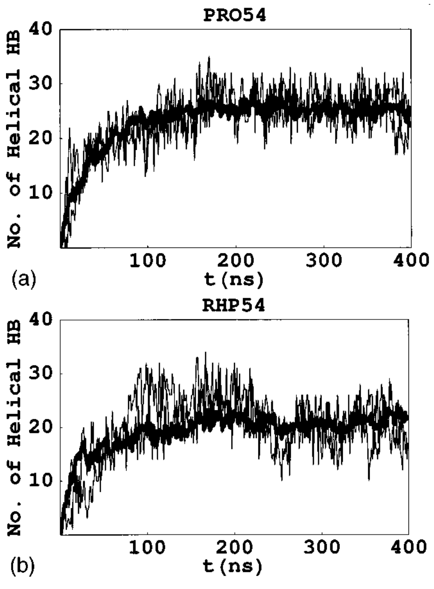

FIG. 10. Number of formed hydrogen bond as a function of time. Two individual cases are drawn with thin curves, the average over all trajectories is in thick curve. (a) PRO54, (b) RHP54.

Figure 11 shows the nativeness Q as a function of time both for PRO54 and RHP54. We see that any native like structure has Q in the range of $0.55 \leqslant Q \leqslant 1$, while RHP54 structures at 300 K typically have Q lower than 0.5. We checked that compact structures that are globally random but have partial helices typically have $Q \sim 0.4$. Fitting of Q averaged over trajectories as a function of time is somewhat difficult. Growth of Q up to about 0.4 is predominantly due to nonspecific compaction, and thus naive fitting leads to a time constant similar to that for collapse. (This may imply that the current definition of Q is not ideal.) Moreover, by inspection, it is obvious that the time for folding is distributed quite widely (50 ns to 0.5 μs) implying very nonexponential relaxation, again making fitting difficult.

D. The designed helical protein (PRO54) and the random heteropolymer (RHP54): Free energy profile

The free energy profile (or surfaces) i.e., potential of mean force, is particularly informative for understanding folding. Many groups have computed such free energy profiles either in one or two dimensions[14,35–38] with different

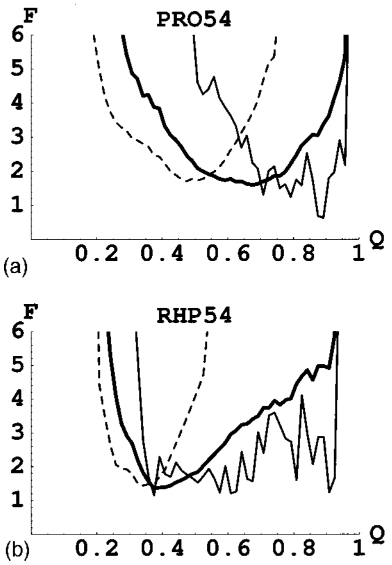

FIG. 12. Free energy profile (potential of mean force) along Q. Thin solid, thick solid, and dashed curves are those at $T=250$ K, 300 K, and 350 K, respectively. (a) PRO54, (b) RHP54.

types of models. With use of the WHAM, we compute such free energy profiles both for PRO54 and RHP54 and compare them.

First, we compute the free energy profile $F(Q)$ with respect to the nativeness Q. Figures 12 depicts the resulting free energy curves for three different temperatures, 250 K (thin solid curve), 300 K (thick solid curve), and 350 K (dashed curve). For PRO54, we see a single minimum for the nativelike state around $Q=0.7$ at 300 K (the thick curve), while there is no other minimum for the globule state. In terms of the Q coordinate used here, despite the nonadditive forces, the PRO54 exhibits downhill folding without any free energy barrier. On the other hand, the free energy profile of the RHP54 has its minimum at $Q=0.4$ that is the typical overlap value between randomly disordered compact states, as mentioned before. At lower temperature (250 K), apparently the free energy curves have more statistical error, though still we can get crude insights. Structures closer to the native structure become still more dominant for PRO54 as expected, while the RHP54 retains a flat region between $Q=0.4$ and $Q=0.9$. Spin glass models predict for the RHP that there are multiple nearly degenerate states that do not resemble each other in structure.[39,40] This is precisely the behavior seen. At low temperature, these states dominate the statistics and give nearly flat curve. At high temperature, the position of the minimum shifts to the left continuously for both cases. Cooperative unfolding upon heating, which is one of the most common feature in real proteins, is not found in RHP54.

We next compute the free energy surface in two dimensions. Figure 13 shows the free energy surface in the (Q,R_g) plane for both PRO54 and RHP54 at 300 K where the red region has lower energy and the blue has high energy. The PRO54 has a relatively narrow region of low energy, while RHP54 distributes somewhat more widely especially in R_g. Because of many frustrated interactions, the hydrophobic core is weaker for RHP54 and the peptide thus easily expands upon heating.

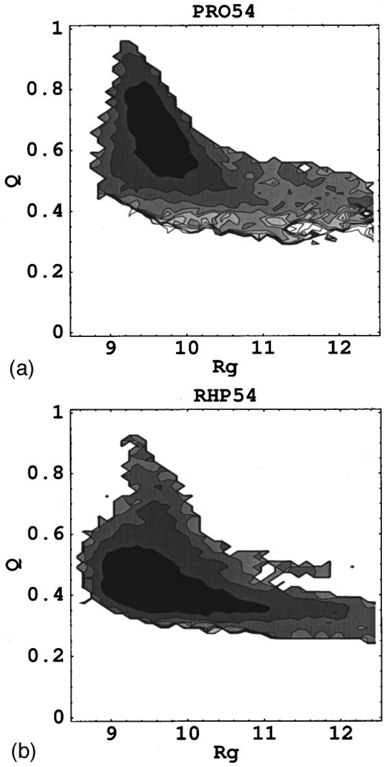

FIG. 13. Free energy surface on Q and R_g plane at $T=300$ K. Energy interval is 1.0 kcal mol^{-1}. Dark region has lower energy. (a) PRO54, (b) RHP54.

The folding temperature T_F can be monitored by looking at the heat capacity as a function of temperature. We present it in Fig. 14 both for PRO54 (solid curve) and RHP54 (dashed curve). The heat capacity is calculated simply as $C_v=(\langle V^2\rangle-\langle V\rangle^2)/(k_BT^2)$, where V is the total potential energy. Both curves exhibit a broad peak around $T\sim305$ K, which corresponds to T_F for PRO54. The designed sequence has a relatively sharper peak here although it still is not very sharp. In contrast to the barrierless behavior in free energy profiles, the heat capacity data implies that the folding transition is (weakly) first order like. It is interesting that, perhaps because of the limited number of amino acid used, i.e., 3 letters, the RHP possesses weak cooperativity, too.

Another important characteristic temperature is the glass transition temperature. Spin glass theory[39] suggests, this is best seen from the probability distribution function of over-

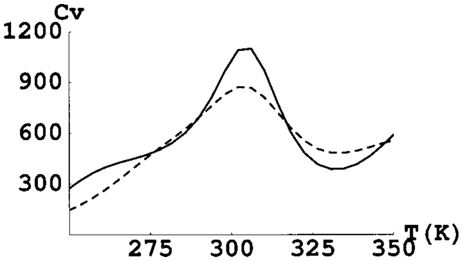

FIG. 14. Heat capacity as a function of temperature. Solid and dashed curves are for PRO54 and RHP54, respectively.

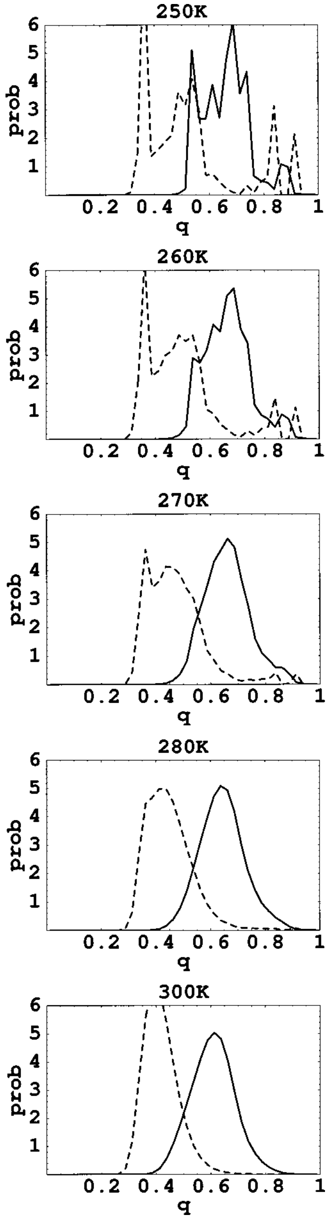

FIG. 15. Probability distribution of mutual overlap q between two structures. Solid and dashed curves are for PRO54 and RHP54, respectively. Temperatures used are 250 K, 260 K, 270 K, 280 K, and 300 K.

lap between two structures sampled in the thermal equilibrium, $P(q)$, defined in the following way. First, we define the overlap between two structures $q_{\nu\mu}$ which resembles to the nativeness Q [Eq. (26)], but instead of the reference to the native structure, computing overlap between the ν and μ structures. Then, the probability distribution is

$$P(q)=\sum_{\nu<\mu} P_{\nu}P_{\mu}\delta(q-q_{\nu\mu}),$$

where P_{ν} is the Boltzmann weight for the structure ν. With use of the trajectories data at 300 K, we apply here the single histogram method (SH) to computing $P(q)$ at several temperatures below 300 K (Fig. 15). At 250 K, RHP54 roughly shows bimodal behavior having peaks around 0.4 and 0.9, that indicates a glassy landscape. The region $q\sim0.4$ corresponds to the overlap between two random collapsed structures, while $q\sim0.9$ represents that two structures are in the same basin in the energy landscape. *Fine structure* in each region may be a signal of a hierarchical landscape, though this could also be an artifact of insufficient sampling. Upon heating, this bimodal distribution merges to a broad unimodal distribution at 280 K, where the protein motion is considered as nonglassy. The onset of bimodal peak found around $T\sim270$ K is assigned to the glass transition temperature. We close this section by remarking that the details of the glassy behavior are very difficult to address numerically, because of nonergodicity. Thus, the current results could have some significant sampling errors.

IV. DISCUSSION

The local density dependent hydrogen bond model shows quite intriguing behavior. First, the stability of a helix depends on its environment as stressed above. A 16 residue segment, as an isolated peptide, does not form a stable helix, but, when it is embedded in the 54 residue protein, it does. This seems to be quite reasonable from a biochemical perspective. Second, in our model, helix formation and compaction are quite concerted at the very beginning of the folding. This is a consequence of the density dependent hydrogen bond model and is consistent with chemical considerations. Without secondary structures, segments of real proteins cannot come together because of large dipoles in the amide proton and carbonyl oxygen. For longer proteins, local compaction instead of a complete global collapse may be sufficient for helix formation.

As mentioned in the previous section, the fluctuations of the model protein in our nativelike structure are still quite large at T_F. This involves a noticeably large change in relative angles between helices and fluctuations of the terminal residues are especially very large. All of these suggest that our nativelike states may correspond to partially ordered molten globule states that have roughly the right topology, but where side chain close packing has not been accomplished. This observation is indeed consistent with the DeGrado group's experimental results who used quite a reduced coding to design a four helix bundle protein.[28] DeGrado *et al.* found that their *de novo* designed 4-helix bundle protein in the nativelike phase fluctuates considerably more than natural highly evolved proteins. This residual degeneracy are seen in some other reduced model simulations[13] as well as a full atom simulation.[11] DeGrado suggested that because of fewer types of amino acids used for the design, packing of the side chains cannot be as tight as in natural proteins. We also use only three kinds of amino acids. Only one type of amino acid is used in the core. Therefore, it is quite natural that our model exhibits quite large fluctuations. Probably more types of amino acids are needed for the tightly packed and less fluctuating native states of most proteins with defined crystal structures.

Related to this point, we have discussed the cooperativity in the folding transition. In the previous section we have shown that the model does not create a substantial barrier between the native state and the collapsed state and thus the folding transition is not very strongly first order (all-or-none) here, which is apparently different from many experiments.

Most simulation studies so far share this feature to varying degrees. On the other hand, the heat capacity as a function of temperature exhibits a broad peak indicating the weak cooperativity. In the current model protein, only one type of amino acid is used in the hydrophobic core and such a homogeneous model cannot have a nondegenerate native minimum often found in real proteins. Both inhomogeneity in the size of the side chain and that in the HP interaction strength may induce a more significantly cooperative transition. This is one of the most probable reasons for missing the barrier. Effects of side chain entropy may need to be explicitly modeled as another type of nonadditive force, too. The values of potential parameters for length scales may be somewhat poor too and thus fine tuning of them might change the behavior.

In the previous sections, we have estimated various time scales involved in folding of the model protein. The time scales that result are smaller than those inferred from experiments. For example, helix formation time is experimentally measured as about 200 ns, whereas we obtained 40 ns. There can be several possibilities for this deviation. First, it is possible that helix formation depends on the detailed sequence of the helical segment as well as tertiary contacts around it and thus the time scales for helix formation are indeed different for different proteins. We note that the time scale for forming helix is ~15 ns in HLX16 but about ~40 ns in PRO54. Another possible reason for this is that we use radius of side chain for ALA (and SER) for most amino acids. Real proteins have more bulky amino acids. With a larger radius, peptide motions should be slower decreasing the discrepancy from experiments. Also the model hydrogen bond may be a little bit too strong making helices too stable. Also we note that using the overdamped Langevin dynamics without any hydrodynamic interaction effects might not be appropriate even for such a slow process because of the large local barriers in the Ramachandran plots which may make transitions underdamped. As for the difference of overall folding times between experiment and the current simulation this may indicate that the simulation addresses only the formation of a structured molten globule state instead of a completely native state. Packing of side chains in the dense globule may take additional time and thus may be rate limiting. All these questions call for further analysis.

It is very interesting that we found qualitatively very different folding trajectories for the designed helical protein. An individual protein falls downhill to the native state and does so through diverse routes. Most current experiments measure properties of an ensemble. Especially the transition state of the averaged folding pathway has been *measured* via the so-called ϕ analysis developed by Fersht *et al.*[5] We recently have developed a statistical mechanical approach to the fast folding free energy surface to compute the averaged folding route and compared it with ϕ-experiment.[41] But experiments have not directly addressed the distribution of folding routes yet. Filling the gap between individual trajectories and experimentally available ensemble averages seems to be a very important and interesting goal.[42]

V. CONCLUSION

We proposed a reduced model of proteins that captures the backbone structure quite completely, while the side chain is simplified to be spheres. The solvent effects are indirectly incorporated without explicitly using water molecules. The interactions are all consistent with physicochemical knowledge and therefore involve nonadditive forces taking into account the effect of solvent. With this model we simulated folding of a designed three helix bundle protein from random conformation for 1 μs, as well as dynamics of a random heteropolymer and the helix–coil transformation of a short peptide. Without introducing a bias-potential (sometimes called a Gō term) to the native structure, the model spontaneously finds nativelike structures for the designed sequence within 0.5 μs. In contrast, the random heteropolymer collapses to a set of nonspecific compact structures with many helical segments, which on further cooling become glassy. Although some fine tuning is necessary, the model looks promising for simulating realistic fast folding of relatively short proteins up to an order of the millisecond time scale.

ACKNOWLEDGMENTS

Support by the JSPS both through Fellowships for Research Abroad and through Research Fellowships for Young Scientists made S.T. possible to stay at the University of Illinois where the study started. P.G.W. is supported by NIH Grant No. PHS 1 R01 GM44557.

[1] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).

[2] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins: Struct., Funct., Genet. **21**, 167 (1995).

[3] P. G. Wolynes, Z. Luthey-Schulten, and J. N. Onuchic, Chem. Biol. **3**, 425 (1996).

[4] B. Nölting, R. Golbik, and A. R. Fersht, Proc. Natl. Acad. Sci. USA **92**, 10668 (1995).

[5] L. A. Itzhaki, D. E. Otzen, and A. R. Fersht, J. Mol. Biol. **254**, 260 (1995).

[6] R. M. Ballew, J. Sabelko, and M. Gruebele, Proc. Natl. Acad. Sci. USA **93**, 5759 (1996).

[7] R. E. Burton, G. S. Huang, M. A. Daugherty, P. W. Fullbright, and T. G. Oas, J. Mol. Biol. **263**, 1996 (311).

[8] V. Muñoz, P. A. Thompson, J. Hofrichter, and W. A. Eaton, Nature (London) **390**, 196 (1997).

[9] W. A. Eaton, P. A. Thompson, C. Chan, S. J. Hagen, and J. Hofrichter, Structure **4**, 1996 (1133).

[10] S. J. Hagen, J. Hofrichter, A. Szabo, and W. A. Eaton, Proc. Natl. Acad. Sci. USA **93**, 11615 (1996).

[11] Y. Duan and P. A. Kollman, Science **282**, 740 (1998).

[12] A. Godzik, A. Kolinski, and J. Skolnick, J. Comput. Chem. **14**, 1194 (1993).

[13] H. Nymeyer, A. E. García, and J. N. Onuchic, Proc. Natl. Acad. Sci. USA **95**, 5921 (1998).

[14] G. F. Berriz, A. M. Gutin, and E. I. Shakhnovich, J. Chem. Phys. **106**, 9276 (1997).

[15] B. M. Pettitt and M. Karplus, Chem. Phys. Lett. **121**, 194 (1985).

[16] M. Pellegrini, N. Grønbech-Jensen, and S. Doniach, J. Chem. Phys. **104**, 8639 (1996).

[17] D. Eisenberg and A. D. McLachlan, Nature (London) **319**, 199 (1986).

[18] N. Gō, Annu. Rev. Biophys. Bioeng. **12**, 183 1983; H. Abe and N. Gō, Biopolymers **20**, 1013 (1980).

[19] N. D. Socci, J. N. Onuchic, and P. G. Wolynes, J. Chem. Phys. **104**, 5860 (1996).

[20] D. Thirumalai and Z. Guo, Biopolymers **35**, 137 (1995); Z. Guo and D. Thirumalai, *ibid.* **36**, 83 (1995).

[21] M. Sasai, Proc. Natl. Acad. Sci. USA **92**, 8438 (1995).

[22] C. Hardin, Z. A. Luthey-Schulten, and P. G. Wolynes, Proteins: Struct. Func., Genet. **34**, 281 (1999).

[23] B. Honig and A. Yang, Adv. Protein Chem. **46**, 27 (1995).

[24] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, J. Comput. Phys. **23**, 327 (1977).

[25] H. A. Scheraga, Macromolecules **10**, 1 (1977).

[26] For example, CHARMM, B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).

[27] Z. Guo and D. Thirumalai, J. Mol. Biol. **263**, 323 (1996).

[28] L. Regan and W. F. DeGrado, Science **241**, 976 (1988).

[29] See, for example, D. E. Smith and A. D. J. Haymet, J. Chem. Phys. **98**, 6445 (1993).

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

[31] W. Kabsch and C. Sander, Biopolymers **22**, 2577 (1983).

[32] P. J. Kraulis, Molscript, J. Appl. Crystallogr. **24**, 946 (1991).

[33] D. S. Riddle, J. V. Santiago, S. T. Bray-Hall, N. Doshi, V. P. Grantcharova, Q. Yi, and D. Baker, Nature, Struct. Biol. **4**, 805 (1997); P. G. Wolynes, *ibid.* , 871 (1997).

[34] D. Hamada, S. Segewa, and Y. Goto, Nature, Struct. Biol. **3**, 868 (1996).

[35] E. M. Boczko and C. L. Brooks III, Science **269**, 393 (1995).

[36] Z. Guo, C. L. Brooks III, and E. M. Boczko, Proc. Natl. Acad. Sci. USA **94**, 10161 (1997).

[37] U. H. E. Hansmann, Y. Okamoto, and J. N. Onuchic, Proteins: Struct. Func., Genet. **34**, 472 (1999).

[38] M. Schaefer, C. Bartels, and M. Karplus, J. Mol. Biol. **284**, 835 (1998).

[39] M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).

[40] S. Takada and P. G. Wolynes, Phys. Rev. E **55**, 4562 (1997).

[41] J. J. Portman, S. Takada, and P. G. Wolynes, Phys. Rev. Lett. **81**, 5237 (1998).

[42] J. N. Onuchic, J. Wang, and P. G. Wolynes, J. Chem. Phys. (in press).