Microscopic theory of protein folding rates. I. Fine structure of the free energy profile and folding routes from a variational approach

John J. Portman[a)]
Department of Physics, University of Illinois, Urbana, Illinois 61801

Shoji Takada
Department of Chemistry, Kobe University, Kobe 657-8501, Japan

Peter G. Wolynes[b)]
Department of Chemistry, University of Illinois, Urbana, Illinois 61801

(Received 31 August 2000; accepted 30 October 2000)

A microscopic theory of the free energy barriers and folding routes for minimally frustrated proteins is presented, greatly expanding on the presentation of the variational approach outlined previously [J. J. Portman, S. Takada, and P. G. Wolynes, Phys. Rev. Lett. **81**, 5237 (1998)]. We choose the λ-repressor protein as an illustrative example and focus on how the polymer chain statistics influence free energy profiles and partially ordered ensembles of structures. In particular, we investigate the role of chain stiffness on the free energy profile and folding routes. We evaluate the applicability of simpler approximations in which the conformations of the protein molecule along the folding route are restricted to have residues that are either entirely folded or unfolded in contiguous stretches. We find that the folding routes obtained from only one contiguous folded region corresponds to a chain with a much greater persistence length than appropriate for natural protein chains, while the folding route obtained from two contiguous folded regions is able to capture the relatively folded regions calculated within the variational approach. The free energy profiles obtained from the contiguous sequence approximations have larger barriers than the more microscopic variational theory which is understood as a consequence of partial ordering. [DOI: 10.1063/1.1334662]

I. INTRODUCTION

Considerable progress has been made in describing protein folding using equilibrium and nonequilibrium statistical mechanics, but a complete formal microscopic kinetic theory has only been sketched. The primary novel features of the modern theory of folding revolve around two somewhat different themes[1–3]—the glassy dynamics expected for most heteropolymers whose sequence is chosen at random,[2–6] and the organized dynamics expected for proteins selected by evolution to fold quickly on a funneled energy landscape.[7–12] The establishment, through selection, of a funneled landscape entails minimizing frustration[1]—the conflict between different energy contributions in a random sequence. Once a funneled landscape is established by selection, however, it is the interplay between entropy and guiding energies of the funnel that figure most prominently in determining the observed kinetics. If we neglect sidechain and solvent degrees of freedom, the entropy depends crucially on the polymer physics of the protein chain. The statistical mechanical theory of both the stable states and some transition states has already been outlined.[13–16] In this paper and its companion[17] we show how the existing framework can be extended to yield a complete microscopic theory of folding rates of completely minimally frustrated proteins. Microscopic calculations of transition state ensembles, activation free energies, and dynamical prefactors involving chain motions can be obtained. A brief report on the early progress of this work has already appeared,[16] but here we fill in the details and also explore some interesting polymer physics issues that we only touched on previously.

The folding transition can be considered as a finite-size phase transition involving two or more stable phases: one a high entropy denatured state with little structure, although perhaps collapsed; and the other the low entropy folded state.[18] A quantitative way to distinguish among these phase is through the magnitude of the fluctuations of each monomer about its average position. In the denatured states, the protein explores many conformations, and because there is little well-defined structure, the fluctuations of a residue about any particular position are relatively large. In the folded state, the conformations are much more restricted and can be described as relatively small fluctuations about the localized positions of the average native structure.[19] These small amplitude deviations are reflected in part by the ''temperature factors'' (or Debye–Waller factors) which can be experimentally obtained from fitting x-ray crystallography data to a model structure that allows for these fluctuations.

The qualitative difference between the liquid and solid

[a)]Current address: Theoretical Biology and Biophysics Group, T-10 MSK710, Los Alamos National Laboratory, Los Alamos, New Mexico 87545.
[b)]Current address: Department of Chemistry and Biochemistry, University of California at San Diego, La Jolla, California 92093.

phases in an ordinary first order crystallization transition can also be described by fluctuation magnitudes reflecting the degree of localization about mean positions. The thermodynamics and to some extent the kinetics of first order phase transitions can be studied using a free energy density functional.[20] To analyze the kinetics of the transition one introduces an approximate density profile that is able to interpolate between the two phases, modeling the formation of crystallites or droplets. The stable phases which are minima of the functional and the critical nucleus which is a saddle point of the functional can then be described through the variational parameters of the trial density. From the point of view of structure, the folded state of the protein is similar to one specific minimum of an amorphous solid in so far as it is not infinitely periodic. One major difference between most inorganic solids and proteins is that proteins are polymers where topology is important; the residues not only have a chemical identity but also a definite sequence. Accordingly, the uniqueness of the folded structure refers not only to the three-dimensional shape, but also to the specific residues localized at the coordinates of the native structure. This suggests that in order to apply the free energy functional formalism to study the folding of a particular protein, one should use order parameters that are local in sequence.

An approach to folding kinetics based on a ''site-resolved'' protein folding free energy functional was presented in Ref. 16. In this approach, a variational free energy surface is introduced directly through a reference Hamiltonian, which provides a good way of approximating the density. The resulting free energy surface is parametrized by the fluctuation magnitudes of each residue about the average, folded conformation. When put into practice the scheme is very similar in spirit to the density functional calculations described above. The variational approach has also been used to characterize the protein folding phase diagram[13,14] as well as to study folding nucleation[15] without attending to the specifics of a given native protein structure. In order to explore the role of the polymer chain statistics in folding, we specialize our calculation to study a particular protein that folds to a known structure. The example we choose, the λ-repressor protein, has been much studied in the laboratory[21,22] but we will explore how its folding routes and free energy profile are changed upon varying polymer statistics sometimes using ''unphysical'' values of the backbone parameters in order to gain insight.

II. GAUSSIAN MODEL FOR A STIFF CHAIN

Consider the Gaussian approximation to the probability density for the n monomer positions $\{\mathbf{r}_i\}$ of a polymer chain

$$\Psi[\{\mathbf{r}\}]\sim\exp\left[-\frac{3}{2a^2}\sum_{ij}\mathbf{r}_i\cdot\Gamma_{ij}\cdot\mathbf{r}_j\right], \tag{1}$$

where a is a microscopic length scale taken to be the mean square distance between adjacent monomers, and we have assumed the mean position vanishes. The correlations of monomer positions in Eq. (1) are given by $\langle\mathbf{r}_i\cdot\mathbf{r}_j\rangle/a^2=[\Gamma^{-1}]_{ij}$.

Different choices of the correlations result in different Gaussian models for the polymer backbone. It is most natural to define the model chain in terms of the correlations between the $n-1$ bond vectors: $\mathbf{a}_i=\mathbf{r}_{i+1}-\mathbf{r}_i$. Denoting correlations between bond vectors by $\langle\mathbf{a}_i\cdot\mathbf{a}_j\rangle/a^2=[\Gamma^{(b)-1}]_{ij}$, the bond correlations and the monomer position correlations are related by

$$\Gamma=M^T\Gamma^{(b)}M, \tag{2}$$

where M is the $(n-1)\times(n)$ nearest neighbor difference matrix

$$M=\begin{bmatrix}-1 & 1 & & \cdots & 0\\ & -1 & 1 & & \vdots\\ \vdots & & \ddots & \ddots & \\ 0 & \cdots & & -1 & 1\end{bmatrix}. \tag{3}$$

A simple way to account for chain stiffness is to assume there is a fixed angle θ between adjacent bonds. This freely rotating chain model[23] has monomer correlations that decay exponentially as $\langle\mathbf{a}_i\cdot\mathbf{a}_{i+l}\rangle/a^2=g^l$, where $g=\cos\theta$. Inverting the matrix of bond correlations and using Eq. (2) gives the inverse of the monomer correlations[24]

$$\Gamma=\frac{1-g}{1+g}K^R+\frac{g}{1-g^2}[K^R]^2-\frac{g^2}{1-g^2}\Delta, \tag{4}$$

where K^R is the Rouse matrix for a nearest neighbor harmonic chain

$$K^R=\begin{bmatrix}1 & -1 & & \cdots & 0\\ -1 & 2 & -1 & & \vdots\\ & \ddots & \ddots & \ddots & \\ \vdots & & -1 & 2 & -1\\ 0 & \cdots & & -1 & 1\end{bmatrix}, \tag{5}$$

and Δ is accounts for the ''boundaries'' at the end of the chain

$$\Delta=\left[\begin{array}{cc|cc}1 & -1 & \cdots & 0\\ -1 & 1 & & \vdots\\ \hline \vdots & & 1 & -1\\ 0 & \cdots & -1 & 1\end{array}\right]. \tag{6}$$

In the continuum limit, this Gaussian polymer model gives the familiar form of wormlike chain that restricts the mean separation between adjacent monomers as well as the local curvature of the chain (e.g., see Ref. 25 and references therein). In this paper, we use the discrete representation and identify the monomers to be the α carbons composing the polypeptide backbone. With this choice, the root mean square separation distance a is the typical distance between adjacent α carbons $a\approx3.8$ Å.

The other parameter in the chain model defines the chain stiffness: as $g\to0$, $\Gamma=K^R$ gives the familiar correlations of a flexible chain ($\langle\mathbf{a}_i\cdot\mathbf{a}_j\rangle=a^2\delta_{ij}$), and as $g\to1$, the correlations correspond to that of a rigid rod ($\langle\mathbf{a}_i\cdot\mathbf{a}_j\rangle=a^2$). Another measure of the chain stiffness is given by the persistence length $l_i=\Sigma_j\langle\mathbf{a}_i\cdot\mathbf{a}_j\rangle/a$.[23] It is possible to introduce nonuniform stiffness parameters (and hence local persistence

lengths) in order to model the different flexibilities of the monomers composing a heteropolymer.[26] For example, the bond correlations can be extracted from the rotational isomeric states model of Flory,[23] or from equilibrium simulations of a detailed polymeric potential.[26–30] Still more complex models can be used to model explicit chiral helical tendency[31] through anisotropic Gaussian correlations. For simplicity, we assume here that the chain stiffness is uniform in this paper, so that the persistence length is related to the chain stiffness by $l \approx a/(1-g)$. For proteins, a reasonable value for the chain stiffness is $g=0.8$ which corresponds to the persistence length of polyalanine, $l=5a\approx 20$ Å.[23,32] In this paper we will also compare chains of other uniform stiffnesses as well.

In addition to the stiff chain Gaussian correlations given by Eq. (1) and Eq. (4), we also include a confining potential that controls the overall size of the polymer chain, which is to say it determines the proximity to the chain collapse transition. To model a collapsed stiff chain, we consider the chain Hamiltonian

$$\beta H_{\text{chain}} = \frac{3}{2a^2} \sum_{ij} \mathbf{r}_i \cdot \Gamma_{ij} \cdot \mathbf{r}_j + \frac{3}{2a^2} B \sum_i \mathbf{r}_i^2, \tag{7}$$

where the parameter B in the second term is conjugate to the radius of gyration of the chain. This confining potential has been used in the reference Hamiltonian in Refs. 13 and 14 where B is treated as a variational parameter. In this paper, B is fixed to give a reasonable radius of gyration of the unfolded state. To establish notation for future use, we rewrite Eq. (7) as

$$\beta H_{\text{chain}} = \frac{3}{2a^2} \sum_{ij} \mathbf{r}_i \cdot [\Gamma^{(\text{ch})}]_{ij} \cdot \mathbf{r}_j, \tag{8}$$

where

$$\Gamma^{(\text{ch})}_{ij} = \Gamma_{ij} + B\delta_{ij} \tag{9}$$

are the inverse monomer correlations of the collapsed stiff chain.

III. VARIATIONAL FREE ENERGY SURFACE

The Hamiltonian for our protein model is

$$H = H_{\text{chain}} + H_{\text{int}}, \tag{10}$$

where H_{chain} is the backbone potential defining the polymeric correlations given in Eq. (8) and H_{int} is the (two-body) interaction potential between distant monomers. The interaction between distant monomers is modeled by a pair potential $u(r)$,

$$H_{\text{int}} = \sum_{[ij]} \epsilon_{ij} u(|\mathbf{r}_i - \mathbf{r}_j|), \tag{11}$$

where ϵ_{ij}, the strength of the interaction, depends on the identity of the residues i and j. The spatial dependence of the interactions between distant monomers consists of an attractive well and a repulsive core. For computational convenience, we approximate the interaction potential as the sum of three Gaussians:

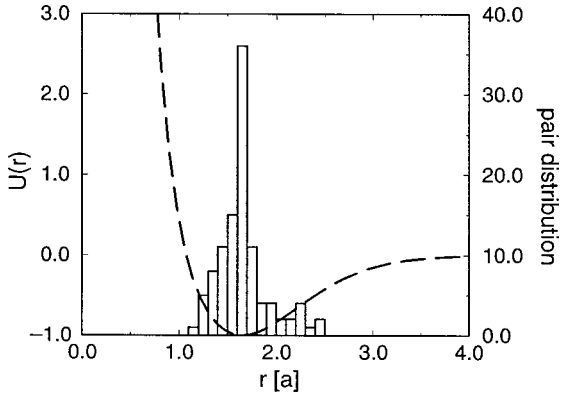

FIG. 1. Interaction potential and λ-repressor native contact distribution. The intermediate- and long-range parameters are given in the text.

$$u(r) = \sum_{k=(s,i,l)} \gamma_k \exp\left[-\frac{3}{2a^2} \alpha_k r^2 \right], \tag{12}$$

where $\alpha_s > \alpha_i > \alpha_l$ are the widths of the short-, intermediate-, and long-range interactions, respectively. The intermediate-range term is repulsive ($\gamma_i > 0$) and the long-range term is attractive ($\gamma_l < 0$); the intermediate- and long-ranged potential parameters are chosen so that the sum of these two terms gives a potential well at an appropriate distance for contacts in the native structure. The short-range term is repulsive ($\gamma_s > 0$) and represents the hard core repulsion between residues. [For an example of $u(r)$ see Fig. 1.] The form of the interaction potential given in Eq. (12) facilitates an analytic expression for the Gaussian average of $u(r)$, though other potentials could also be considered within the framework of these calculations.

We approximate the free energy surface of the protein using a reference Hamiltonian that corresponds to a polymer in a nonuniform external field that constrains the monomers to lie near their locations in the native structure $\{\mathbf{r}_i^N\}$

$$\beta H_0 = \beta H_{\text{chain}} + \frac{3}{2a^2} \sum C_i (\mathbf{r}_i - \mathbf{r}_i^N)^2. \tag{13}$$

The strengths of the harmonic constraints, $\{C_i\}$, are conjugate to the fluctuations of the polymer about each of the native positions. The external constraints in H_0 influence both the correlations G_{ij} and average positions $\{\mathbf{s}_i\}$ of the monomers composing the reference chain

$$G_{ij} \equiv \langle \delta\mathbf{r}_i \cdot \delta\mathbf{r}_j \rangle_0 / a^2 = [\Gamma^{(0)}]^{-1}_{ij}, \tag{14}$$

$$\mathbf{s}_i \equiv \langle \mathbf{r}_i \rangle_0 = \sum_j G_{ij} C_j \mathbf{r}_j^N, \tag{15}$$

where $\delta\mathbf{r}_i$ is the position of the ith monomer relative to the average

$$\delta\mathbf{r}_i = \mathbf{r}_i - \langle \mathbf{r}_i \rangle_0 = \mathbf{r}_i - \mathbf{s}_i, \tag{16}$$

and $\Gamma^{(0)}$ is the matrix of coefficients of the quadratic terms of H_0

$$\Gamma^{(0)}_{ij} = \Gamma^{(\text{ch})}_{ij} + C_i \delta_{ij}. \tag{17}$$

From the magnitude of these fluctuations, this reference Hamiltonian can distinguish the two stable phases of the pro-

tein (as described above): the globule corresponds to large fluctuations (weak constraints) and folded states correspond small fluctuations (strong constraints).

We consider the variational free energy surface parametrized by the constraint parameters $\{C_i\}$

$$F[\{C\}] = -k_B T \log Z_0 + \langle H - H_0 \rangle_0, \tag{18}$$

where $Z_0 = \mathrm{Tr}[e^{-\beta H_0}]$ is the partition function of the reference Hamiltonian, and $\langle \cdots \rangle_0 = \mathrm{Tr}[\cdots e^{-\beta H_0}]/Z_0$ denotes the average taken with respect to H_0. Substituting the expressions for H and H_0 gives the variational free energy $F = E - ST$, where

$$E[\{C\}] = \sum_{[ij]} \epsilon_{ij} \langle u(|\mathbf{r}_i - \mathbf{r}_j|) \rangle_0 \tag{19}$$

and

$$S[\{C\}]/k_B = \log Z_0 + \frac{3}{2a^2} \sum C_i \langle (\mathbf{r}_i - \mathbf{r}_i^N)^2 \rangle_0 \tag{20}$$

are the expressions for the energy $E[\{C\}]$ and entropy $S[\{C\}]$ as functions of the variational constraints.

The averages over H_0 in Eq. (19) and Eq. (20) can be expressed in terms of G and $\{\mathbf{s}_i\}$, because $e^{-\beta H_0}$ is a Gaussian distribution. One instructive way to calculate the averages is to introduce approximations to the density of monomer i, $\rho_i^1(\mathbf{r}) = \langle \delta(\mathbf{r} - \mathbf{r}_i) \rangle_0$,

$$\rho_i^1(\mathbf{r}) = \left[\frac{3}{2\pi a^2 G_{ii}} \right]^{3/2} \exp\left[-\frac{3(\mathbf{r} - \mathbf{s}_i)^2}{2a^2 G_{ii}} \right], \tag{21}$$

and the pair density between i and j, $\rho_{ij}^2(\mathbf{r}) = \langle \delta(\mathbf{r} - (\mathbf{r}_i - \mathbf{r}_j)) \rangle_0$,

$$\rho_{ij}^2(\mathbf{r}) = \left[\frac{3}{2\pi a^2 \delta G_{ij}} \right]^{3/2} \exp\left[-\frac{3(\mathbf{r} - (\mathbf{s}_i - \mathbf{s}_j))^2}{2a^2 \delta G_{ij}} \right], \tag{22}$$

where $\delta G_{ij} = \langle (\delta \mathbf{r}_i - \delta \mathbf{r}_j)^2 \rangle_0 / a^2 = G_{ii} + G_{jj} - 2G_{ij}$. These densities depend on the constraint parameters $\{C_i\}$ through G and $\{\mathbf{s}_i\}$. Averages over H_0 can be calculated through $\rho_i^1(\mathbf{r})$ and $\rho_{ij}^2(\mathbf{r})$, for example,

$$\langle u(|\mathbf{r}_i - \mathbf{r}_j|) \rangle_0 = \int d\mathbf{r}\, \rho_{ij}^2(\mathbf{r}) u(r). \tag{23}$$

In this way, the variational free energy can be viewed as a density functional with a particular approximation to the density that simultaneously incorporates the polymeric correlations and the monomeric fluctuations about the average positions.

It is straightforward to calculate the entropy and energy in terms of the monomer correlations and mean positions. After some manipulation (see Appendix), the entropy can be written as

$$S[\{C\}] = \frac{3}{2} \log \det G - \frac{3}{2a^2} \sum_{ij} s_i \cdot \Gamma_{ij}^{(\mathrm{ch})} \cdot s_j + \frac{3}{2} \sum_i C_i G_{ii}. \tag{24}$$

This expression can be interpreted as follows: the first term is the entropy of the chain due to polymeric fluctuations, the second term is the entropy loss of fixing the mean positions of each monomer, and the last term is the vibrational entropy about the mean positions $(= (3/2a^2) \Sigma C_i \langle \delta \mathbf{r}_i^2 \rangle_0)$. Similarly, the pair potential can be averaged over H_0 to give the energy

$$E[\{C\}] = \sum_{[ij]} \epsilon_{ij} u_{ij}, \tag{25}$$

where

$$u_{ij} = \langle u(|\mathbf{r}_{ij}|) \rangle_0 = \sum_{k=(s,i,l)} \frac{\gamma_k}{(1 + \alpha_k \delta G_{ij})^{3/2}} \exp\left[-\frac{3}{2a^2} \frac{\alpha_k (\mathbf{s}_i - \mathbf{s}_j)^2}{1 + \alpha_k \delta G_{ij}} \right]. \tag{26}$$

Finally, we choose to measure the free energy relative to the unconstrained chain

$$\Delta F[\{C\}] = \Delta E[\{C\}] - T \Delta S[\{C\}], \tag{27}$$

where, for example, $\Delta F[\{C\}] = F[\{C\}] - F[\{C = 0\}]$.

The reference Hamiltonian plays such a prominent role in the variational theory (and in subsequent calculations of folding dynamics) that it warrants further comment. Not surprisingly, other Gaussian models have been previously introduced to model the polymers with fixed contacts or crosslinks. A Hamiltonian of the form

$$\beta H_{\mathrm{hc}} = \frac{3}{2a^2} \sum (\mathbf{r}_i - \mathbf{r}_{i+1})^2 + C \sum_{[ij]} (\mathbf{r}_i - \mathbf{r}_j)^2 \tag{28}$$

has been used to study the thermodynamics[33–38] and dynamics[39] of polymers crosslinked at the sites specified by the pairs $[ij]$. This harmonic contact Hamiltonian, H_{hc}, has also been used to model the vibrations of folded proteins,[40–42] with the set $[ij]$ limited to the contacts of the native structure, and it was found that the relative magnitudes of monomer fluctuations agree well with measured temperature factors.[40] In contrast to our reference Hamiltonian, H_{hc} is translationally invariant, independent of an explicit native structure. While enforcing this symmetry has some advantages, the potential well is centered at the origin. We note that recently, H_{hc} was used as a mean spherical model for protein folding where a nonlinear constraint was added so that the average monomer positions would lie on the surface of a sphere, preventing the polymer conformation from collapsing to the origin; the minima of H_{hc} with the added imposed condition can yield a meaningful average structure.[43] Nevertheless, the mean locations of the residues are quite distorted. Consequently, H_{hc} is not as well suited as the Hamiltonian we use to describe the protein folding transition where the disordered globule and the structured folded state are separated by a barrier composed of an ensemble of partially ordered conformations. We have, however, investigated H_{hc} as a reference Hamiltonian in the variational context. It gives similar results to those described here.

IV. ORDER PARAMETERS AND FOLDING ROUTES

Setting values for the constraint parameters $\{C\}$ corresponds to selecting an ensemble of conformations specified

by $\{\mathbf{s}\}$ and G. The energy of a pair is most stabilizing when the pair density is contained in the potential well, i.e., the mean separation between monomers is within the well and the fluctuations are relatively small. Accompanying this stabilization, however, is the entropy loss of localizing the positions of the pair. In general, when there are many nonzero constraints, the entropy loss due to localization is given by Eq. (24) through the correlations. In other free energy functionals[44–48] this entropy loss is estimated but there are difficulties in considering only the entropy loss of individual pairs forming loops because the total entropy loss is inherently nonadditive.[49,50] The relationship between the two approaches is much like the difference between the Thomas–Fermi and Hohenberg–Kohn estimates of the kinetic energy in quantum mechanical density functional theory.[51] The values of the constraint parameters corresponding to the local minima of $F[\{C_i\}]$ are a compromise between the energy and entropy decrease of forming contacts. Similarly, the constraints that correspond to saddle point of $F[\{C_i\}]$ also reflect this competition of energy and entropy, because the free energy is minimized in all directions except the unstable mode along which there is a maximum.

The configurational ensemble corresponding to a given set of constraint parameters $\{C_i\}$ can also be described by density like order parameters that depend on the local mean square fluctuations. For any set of functions of the chain positions $\{A_i[\{\mathbf{r}\}]\}$, we can define the order parameter $A_i[\{C\}]=\langle A_i\rangle_0$, as a function of the constraints. This relationship can be inverted locally provided the Jacobian is nonsingular, $\det J\neq 0$ with $J_{ij}=\partial A_i/\partial C_j$. Since these order parameters are a function of the constraints we can then parametrize the free energy by $\tilde{F}[\{A\}]=F[\{C\}]$ with $C_i=C_i[\{A\}]$. For example, the form of H_0 suggests that the local mean square fluctuations (related to Debye–Waller factors)

$$\mathcal{B}_i=\langle \delta\mathbf{r}_i^2\rangle_0=G_{ii}a^2 \tag{29}$$

are a natural set order parameters for the reference Hamiltonian. (Indeed, this is what motivated our choice of H_0.) In studying the dynamics of barrier crossing in the companion paper, it will prove useful to consider a related but different measure of native similarity. With any identification of the order parameters, we can study the properties of the free energy in $\{C_i\}$ space to describe the folding, and then characterize the corresponding ensembles through structural order parameters as equilibrium averages with respect to H_0.

We calculate the transition states involved in the folding by searching for saddle points in $F[\{C_i\}]$ using an eigenvector-following algorithm.[52] This algorithm is similar to Newton's method for optimization, but involves diagonalizing the Hessian matrix, $\partial^2F/\partial C_i\partial C_j$, at each iteration. In this routine, the point is updated by stepping in a direction to maximize along the eigenvector with the lowest eigenvalue and minimize along all others. To find a minimum, a step is taken to minimize along all eigenvectors of the Hessian. In order to use this algorithm, we must be able to differentiate the free energy with respect to $\{C_i\}$, $\partial_\alpha F\equiv\partial F/\partial C_\alpha$. These derivatives can be easily computed by the chain rule using the elementary derivatives $\partial_\alpha G_{ij}=-G_{i\alpha}G_{\alpha j}$ and $\partial_\alpha(\log\det G)=-G_{\alpha\alpha}$. The explicit expressions for the derivatives of the energy and entropy are not given here; they are straightforward to derive, and not very illuminating.

The saddle points and local minima characterize the average folding routes from this theory. These average pathways are found as follows. The globule and native states are identified by the local minima with the largest and smallest entropy, respectively. These are easy to identify, because the globule is the only stable minimum at high temperature and the native state is the only one at low temperature; these minima can be used as the initial guesses for the optimization algorithm for incremental temperature changes until we have these minima at the same temperature. Using linear combinations of these two sets of constraints as initial guesses, we search for a saddle point. From this saddle point, we perturb the set of constraints $\{C_i\}$ along the unstable eigenvector and use the eigenvector following algorithm with a small step size to find the closest minimum. This gives two local minima, one for each direction along the unstable eigenvector, connected by the saddle point. This process is repeated until the globule and native state are connected by a series of local minima and saddle points. We identify this connected sequence as the average folding route, characterizing the transition states and local minima that are important in the folding kinetics. We note that in the example below for the λ-repressor protein, only one folding route was found, but this is not a general result of the theory. The same procedure applied to the SH3 domain has shown that there may be multiple routes.

V. FOLDING ROUTES EXAMPLE: λ-REPRESSOR

In this section, we illustrate the variational theory by studying the folding of a variant of the λ-repressor protein. λ_{6-85}, is a small (80 residue) protein consisting of five helices in the native structure.[53,54] This protein folds extremely rapidly following two-state kinetics.[21] From nuclear magnetic resonance (NMR) measurements of the folding rate of various mutants, Oas and co-workers concluded that the structure of the transition state consisted mainly of residues in helices H1 and H4.[22] In Ref. 16 we compared folding routes calculated from the variational theory with these measurements, using a reasonable choice for the persistence length of the polypeptide chain. We investigate here how these results depend on different values of the persistence length (or chain stiffness). These studies allow us to see how some recent simplified approaches to free energy profiles based on assuming complete contiguous sequence folding[46–48] become more exact as chain stiffness increases.

A. Model parameters

To apply the theory, we need to specify the parameters that describe the interaction potential between residues and the polymer chain characteristics. The parameters of the present paper are the same as those chosen in Ref. 16, but we describe our rationale for this choice in greater detail. For this fast folding protein, we consider a $\overline{\text{Go}}$ model for the interactions. This means that the sum over residues $[ij]$ in Eq. (11) is limited to the set of contacts found in the native

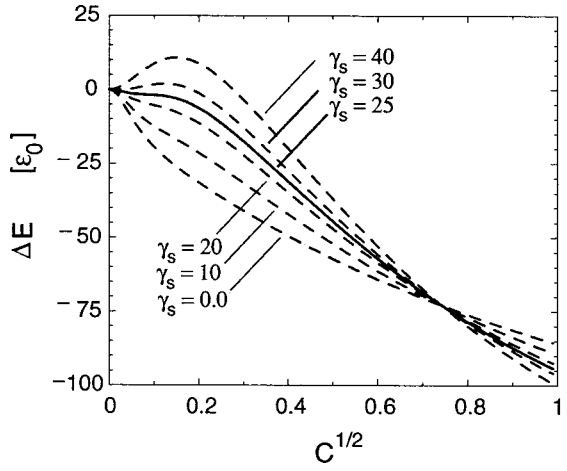

FIG. 2. Energy in units of the Miyazawa–Jernigan energy scale ϵ_0 as a function of the constraint, C. All monomers have equal constraint (one-dimensional), and chain stiffness $g=0.8$. γ_s is the strength of the short-range interaction and the Gaussian width is $\alpha_s=3$. The intermediate- and long-range interaction parameters are the same as in Fig. 1.

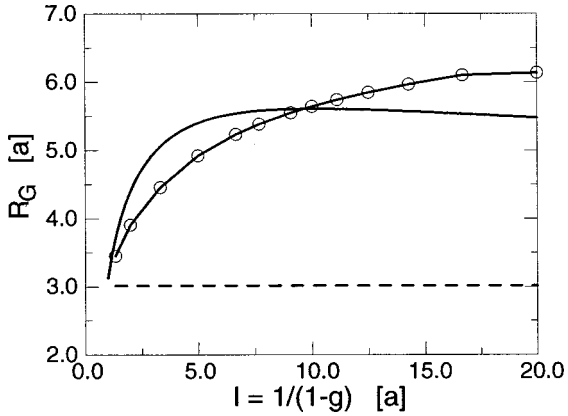

FIG. 3. Radius of gyration as a function of persistence length (in units of monomer spacing a) for λ-repressor ($n=80$) and confinement parameter $B=10^{-3}$: no constraints (solid), globule (○), and native coordinates (dashed).

structure. This set is defined to be pairs of residues ($i+4 \leqslant j$) that have β carbons (α carbons for glycine) distances within a 6.5 Å cutoff in the folded structure. A cutoff between 6–8 Å is commonly used to define contacts in a Gō model, though the precise value is not generally important. We also include in this set residue pairs that are likely to have hydrogen bonds (as determined by the DSSP algorithm[55]) but fall outside the cutoff. The strength of the interactions for this set depends on the residue identities of the pair. We take the well depths ϵ_{ij} to be the magnitude of the Miyazawa–Jernigan energy parameters reported in Ref. 56 in units of $\epsilon_0=k_BT$.

The parameters for the interaction potential $u(r)$ in Eq. (12) are constrained by the $C_\alpha-C_\alpha$ distances of the set of native contacts. The intermediate- and long-range parameters are chosen so that the sum of these Gaussians has an attractive well that contains all the native contacts and has a minimum value $u(r^\star)=-1$ at the most probable contact distance $r^\star=1.6a$. The contact distribution and potential with $(\gamma_i,\alpha_i;\gamma_l,\alpha_l)=(9,0.54;-6,0.27)$ are shown in Fig. 1.

The short-range interactions represent the hard core repulsion between the monomers which controls the density of the collapsed polymer. Due to the Gaussian chain approximation, the pair density given by Eq. (22) has nonzero density at short distances, and hence even a delocalized pair has energy contributions from both the repulsive and attractive components of the potential. In the model, the short-range Gaussian amounts to an effective potential that balances the attractive potential for relatively unconstrained polymers. Choosing the repulsion by this criterion is analogous to finding Θ solvent conditions for the unconstrained polymer (such as the globule). To determine a reasonable repulsive potential for this model, we consider a one-dimensional approximation to the variational free energy by setting all the constraints equal, $C_i=C$. As can be seen in Fig. 2, the energy as a function of C is monotonically stabilizing if γ_s is small, and has a barrier if γ_s large. The parameters for the repulsive potential are chosen so that the energy is relatively constant for small values of the constraint parameter: $(\gamma_s,\alpha_s)=(25,3.0)$. (This particular value of the strength depends on the width of the short-ranged Gaussian which has chosen somewhat arbitrarily.)

The remaining parameter to be specified is B, the strength of the diagonal confinement term. This confinement parameter is a small constant that effects the fluctuations of unconstrained segments of the chain since the constraint parameters are also diagonal terms in the inverse correlation matrix. Figure 3 shows the radius of gyration,

$$R_G^2=1/n^2\sum_{i\leqslant j}\langle(\mathbf{r}_i-\mathbf{r}_j)^2\rangle_0, \tag{30}$$

evaluated at the native coordinates (dashed) and an unconstrained chain $C_i=0$ (solid) as a function of persistence length for $B=10^{-3}$. The unconstrained radius of gyration rises rapidly as the persistence length increases and saturates to a value less than twice the native radius of gyration. The radius of gyration, which in the absence of confinement approaches a large value in the stiff chain limit,[57] is seen to be limited by the confinement term. This tension between local stiffness and confinement is responsible for the plateau in R_G. Figure 3 also shows the radius of gyration evaluated at the constraints corresponding to the globule minimum (○). For persistence lengths less than $l\approx 10a$ the globule constraints lead to a smaller radius of gyration than the free chain value, and for larger persistence length chains the radius of gyrations is somewhat larger. Although it is possible to choose values of B in order to set R_G of the globule for each chain stiffness, we have chosen to illustrate the effects on the polymer conformations by fixing the confinement and independently varying the chain stiffness.

B. Two-dimensional illustration

Because the results of calculations in the full variational space (one constraint parameter for each monomer) are somewhat complicated to present, we will illustrate the model in lower dimensions to give the reader an intuitive feeling for the multidimensional free energy surface and how folding routes are obtained from the model. We consider a

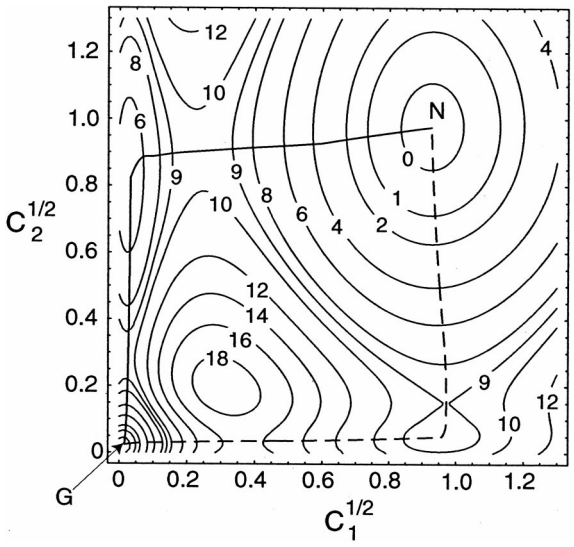

FIG. 4. Contour plot of free energy in units of the Miyazawa–Jernigan energy scale ϵ_0 for the two-dimensional surface (see text). The lines indicate the average folding routes: Path 1 (dotted), Path 2 (solid). The chain stiffness is $g=0.8$.

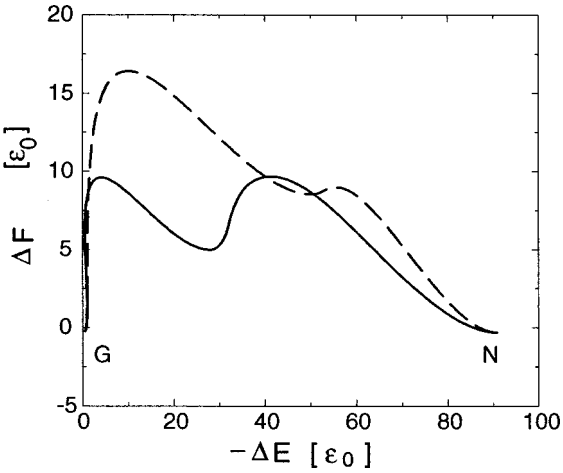

FIG. 5. Free energy profile for both paths indicated in Fig. 4.

two-dimensional approximation in which we group the protein into two segments and assign a variational spring constant to each group. In this example, monomers with index 1–50 correspond to C_1 and the rest correspond to C_2; these two groups correspond roughly to the helices H1–H3 and H4–H5, respectively. This grouping has a loose correspondence to the folding route obtained from the full variational calculations discussed in the next section.

The free energy surface shown in Fig. 4 has two distinct low energy paths determined by the saddle points connecting the globule (G) and native (N) states. The average folding routes, as defined by the path from the saddle points to the local minima, are determined by the eigenvector-following algorithm. As can be seen in Fig. 4 these routes are very close to the steepest descents path. Along the two routes from G to N, in path 1 (dotted line) the constraint C_1 progressively increase followed by the increase of constraint C_2, whereas along path 2 (solid line) this order is reversed with C_2 increasing before C_1.

The free energy of these paths is plotted parametrically versus the energy [$\Delta F(C_1,C_2)$ vs $\Delta E(C_1,C_2)$] in Fig. 5, giving a free energy profile where the saddle points appear as local maxima. Path 2 is the relatively favored route since it has the lower barrier to folding. We note that a rough description of the free energy profile can be represented by connecting the stationary points of $F(C_1,C_2)$ by straight lines.

The ensemble of conformations composing the average folding route can be characterized by the magnitude of the fluctuations of each residue ($\mathcal{B}_i=\langle \delta \mathbf{r}_i^2\rangle_0$) for any set of constraints along these paths. Figure 6 shows the monomer fluctuations evaluated at the constraints corresponding to the globule state and the saddle points of both paths. These fluctuations give a description of folding consistent with the two-dimensional surface parametrized by the constraints: in path 1 residues in helices H1–H3 become structured at TS_1 followed by H4–H5 at TS_2, and in path 2 the order is reversed. For a given set of the constraints (i.e., a given saddle point), the fluctuations are seen to smoothly interpolate between the two groups of monomers. The precise shape of the interface depends on the value of the chain stiffness. A stiffer chain would tend to suppress variations as a function of sequence index resulting in a more gradual interface. As will be seen in the full dimensional calculations, decreasing the chain stiffness allows the magnitude of the fluctuations to change more rapidly between successive monomers, as expected.

C. Fine structure of the free energy profile: Multiple transition states and folding routes

We follow the same analysis outlined in the two-dimensional illustration to describe the folding paths calculated in the full variational space, but the free energy profile is represented by the saddle points and local minima rather than as a continuous path.

In this study, we focus on how the folding routes depend on the persistence length of the chain. To put various parameters into context, the homopolymers polyglycine, polyalanine, and polyproline have persistence lengths[58] $l=6$ Å

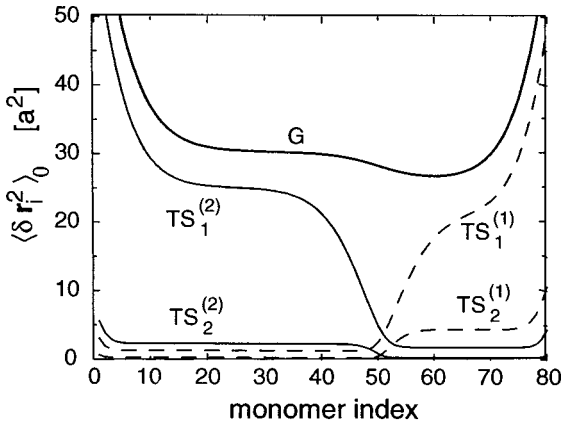

FIG. 6. Fluctuations vs sequence index $\langle \delta \mathbf{r}_i^2\rangle_0 = G_{ii}a$ where a is the distance between successive monomers, evaluated at the saddle points for path 1 (dotted) and path 2 (solid) shown in Fig. 5. Fluctuations of the native (N) and globule (G) are also shown.

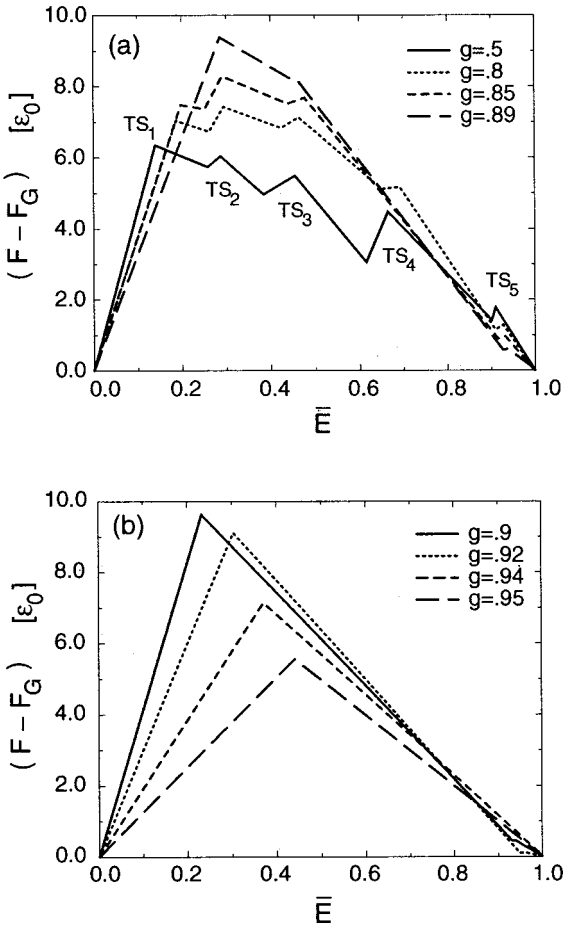

FIG. 7. Free energy profile vs normalized energy [Eq. (31)], for different persistence lengths: (a) $g=0.5$, 0.8, 0.85, 0.89. (b) $g=0.9$, 0.92, 0.94, 0.95.

($\approx 2a$), 20 Å($\approx 5a$), and 220 Å($\approx 60a$), respectively.[23,32] In the freely rotating chain model $l=a/(1-g)$, so that these persistence lengths correspond to the chain stiffness parameters $g=0.5$, 0.8, and 0.98, respectively. Modeling the protein backbone with a single uniform chain stiffness parameter, we take the chain stiffness of polyalanine to be a reasonable value for the protein backbone.

We report the free energy profile at the folding transition temperature, T_f. This is the temperature at which the folded and globule ensembles have the same equilibrium probability ($\Delta F_G=\Delta F_N$). Since the entropy and energy loss in these states depend on the chain stiffness, T_f is different for different persistence lengths. (For a flexible chain $l=2a$, $k_BT_f/\epsilon_0\approx 1.2$, whereas for a very stiff chain $l=20a$, $k_BT_f/\epsilon_0\approx 2.2$.) Similarly, the unconstrained ensemble (which we define as the zero of the free energy) depends on the persistence length. Consequently, to compare the folding profiles of chains with different persistence lengths, it is convenient to plot the free energy profile relative to the globule free energy against a normalized energy coordinate

$$\bar{E}=(\Delta E-\Delta E^G)/(\Delta E^N-\Delta E^G), \tag{31}$$

where ΔE^G and ΔE^N are the globule and native energy changes, respectively. $\bar{E}$ is the fractional stabilization energy and equals 0 at the globule state and 1 at the native state.

Figure 7 shows the free energy profile for chain stiffness parameters ranging from $g=0.5$ to $g=0.95$ ($l=2a$ to $l=20a$). The free energy profile versus energy for a flexible chain ($l=2a$) is shown as the solid curve in Fig. 7(a). The profile exhibits five transition states (saddle points) separated by local minima. We specify the transition states sequentially from the globule minimum (G) to the native minimum (N), and the minima by the index of the adjacent transition state (e.g., the minimum between TS_1 and TS_2 is denoted by min_{12}). The profile can be described as a rugged fine structure resulting from the different structural ensembles of the local minima superimposed on a single average free energy barrier. The fine structure on the profile is modest in magnitude amounting to a stabilization of a ''high energy intermediate'' by at most $1k_BT_f$. This fine structure should not be confused with the ruggedness due to frustration,[1] which we call ''transverse'' ruggedness, coming from degrees of freedom different from the one plotted; instead, the fine structure is better described as ''longitudinal'' ruggedness (along the reaction coordinate).[59] This is a common feature of many models even those which consider only native contacts[12] and which therefore have perfectly funnel-like surfaces.

Starting with the most flexible chains shown in Fig. 7(a), as the stiffness increases the magnitude of the barrier increases and the the minima separating the transition states become relatively still more shallow (the longitudinal ruggedness diminishes). A flexible chain can take advantage of particularly strong contacts while losing a relatively small amount of entropy to localize the pair. This gives rise to the possibility of relatively stable local minima and lower free energy barriers. Stiffer chains must localize larger segments of the chain (on the order of the persistence length) resulting fewer distinct local minima and larger free energy barriers. For the largest chain stiffness in Fig. 7(a) ($l\approx 9a$) the local minima have nearly disappeared, leaving a free energy profile with a single transition state ensemble. The profile for even stiffer chains ($l=10a$ to $l=20a$) is shown in Fig. 7(b). In this parameter range, as the chain stiffness increases further the barrier decreases and the single transition state occurs at a larger fractional stabilization energy (i.e., the energy of transition state becomes closer to the native energy). When the persistence length is large enough, large sections of the chain must be constrained resulting in a more folded transition state. Evidently, for very stiff chains the transition state is similar enough to the native minimum that the barrier decreases.

The barrier height is plotted as a function of persistence length in Fig. 8 (solid line). The maximum barrier height occurs for a persistence length $l\approx 10a$ with an increase of approximately 70% relative to the barrier height for the most flexible chain considered. However, the appropriate energy scale for the transition is k_BT_f. In units of k_BT_f, the barrier height is relatively constant for a wide range of persistence lengths (dashed line), changing only approximately 10% for persistence lengths up to $l\approx 12a$. The robustness of the barrier height at T_f is interesting since the persistence length of real proteins in the laboratory is not precisely known.

Each ensemble of structures along the average folding route can be characterized by the local temperature factors [Eq. (29)]. The fluctuations corresponding to the transition

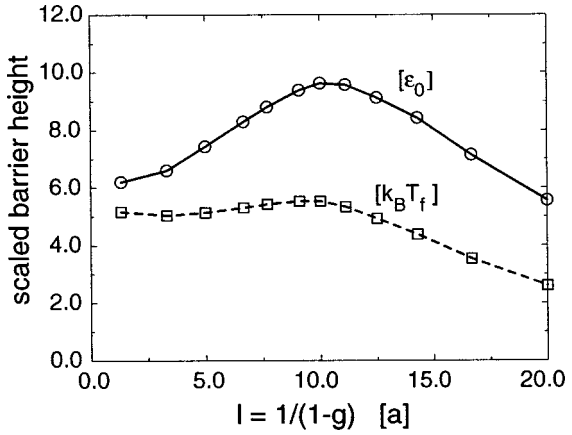

FIG. 8. Barrier height vs persistence length. Free energy scaled by Miyazawa–Jernigan energy scale ϵ_0 (solid), and by the folding temperatures $k_B T_f$ (dashed).

states for a flexible chain ($l=2a$) are shown as the dotted curves in Fig. 9(a). The fluctuations corresponding to the local minima are within those of adjacent saddle points; for clarity, only one local minimum, $\min_{23}$, is shown (solid line). The folding route can be characterized by considering the structure of the transition states from the globule to the native structure. The first transition state ensemble TS_1 is described by the ordering of helices H4–H5, stabilized by the partial localization of a region of helix H1, while residues in helices H2–H3 remain delocalized. In the subsequent transition states, helix H1 becomes progressively more ordered, while helices H2–H3 continue to have large fluctuations and are the last to order. This scenario for the folding of λ-repressor agrees with the interpretation of kinetic data based on ϕ-value analysis.[22]

The temperature factors for a larger persistence length ($l=5a$) are shown in Fig. 9(b). The behavior of the fluctuations describing the transition state ensembles is qualitatively similar to the more flexible chain, though the magnitude of the fluctuations of the disordered regions is larger. Some of the detailed features shown in Fig. 9(a) have been smoothed out since the chain stiffness suppresses variations along the sequence smaller than the persistence length. For example, the very specific localization of a helix H1 residue shown in the $\min_{23}$ curve of the more flexible chain [Fig. 9(a)] has been broadened to a larger region for the stiffer chain. These differences are rather subtle, but the comparison is useful to illustrate the progression to larger chain stiffnesses. The fluctuations for the single transition state characteristic of larger persistence lengths is shown in Fig. 9(c). (Note, here the different curves correspond to different persistence lengths rather than transition states along the same folding route.) While the general shape of the curve is maintained, the magnitude of the fluctuations becomes increasingly more like those of the native state, and in the largest persistence length considered $l=20a$ only the end segments of the chain are significantly disordered.

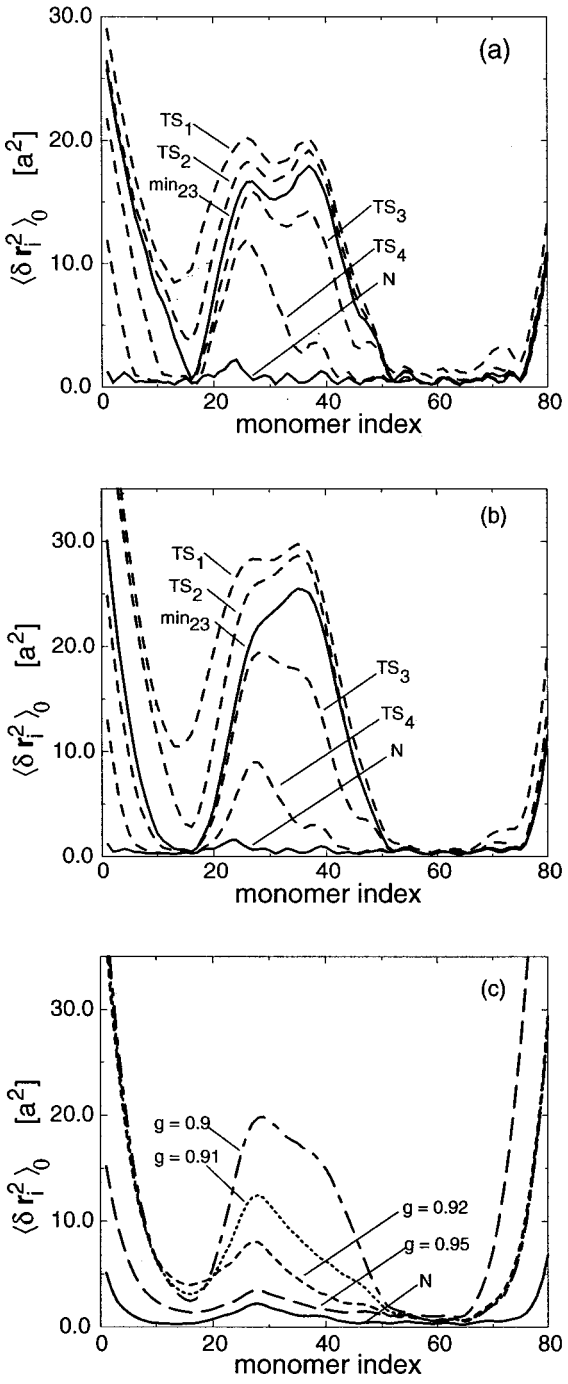

FIG. 9. Fluctuations vs sequence index $\langle \delta \mathbf{r}_i^2 \rangle_0 = G_{ii} a$ where a is the distance between successive monomers, of selected stationary points on the folding route for different persistence lengths. (a) $l=2a$ ($g=0.5$), (b) $l=5a$ ($g=0.8$), (c) $l \approx 10$, 11, 13, 20 ($g=0.9, 0.91, 0.92, 0.95$).

D. Contiguous sequence approximations

Several recent and apparently accurate estimates of folding kinetic parameters have assumed that the transition state ensemble can be described by assigning contiguous segments to be either folded or not and allowing the sequence to be passed into such fully folded or unfolded configurations.[46–48] As Plotkin *et al.* argued, such a contiguous sequence approximation (CSA) should apply to late transition states where the entropy is low.[60] How do the polymer's characteristics determine the quality of this approximation?

Following the contiguous sequence approach, to simplify the problem we reduce the number of states in the variational theory's description of the protein ensembles by

restricting the conformations to only those in which the structured residues are fully native and contiguous in sequence (single CSA), or alternatively, we consider two contiguous stretches of native residues (double CSA).[61] For each fixed value of the number of folded residues, N_f, we find the minimum free energy configuration (satisfying the single or double contiguous constraint). In this way we can construct a free energy profile as a function of N_f. This construction neglects the connectivity of the path since the minimum free energy configuration with N_f folded residues may not be simply related to that with N_f+1. This connectivity is an added complication to the approach and can be treated by the methods presented Refs. 46 and 47. For the purposes of this illustration, we neglect this aspect of the approximation. The approach outlined here is a great simplification of the more complete variational formalism since it avoids the relatively difficult numerical calculation of finding saddle points as a function of degree of ordering. Still, the number of configurations needed in the double CSA is quite large, but tractable ($\approx$1.7 million for λ-repressor). On the other hand, these approximations are rather restrictive since they neglect partial ordering and provide a less microscopic characterization of the folding route.

One issue in comparing the exact variational and the CSA results is how to measure the partial order characterizing the stationary ensembles in a global way, comparable to the number of fully native residues in the CSAs. For a given stationary point, we can compare by estimating N_f through the normalized fluctuations

$$\overline{\mathcal{B}_i}=(\mathcal{B}_i-\mathcal{B}_i^G)/(\mathcal{B}_i^N-\mathcal{B}_i^G), \tag{32}$$

where the superscripts G and N denote the fluctuations evaluated at the globule and native state, respectively. As a rough estimate, we define N_f to be the number of residues with $\overline{\mathcal{B}_i}\geqslant 0.95$.

The free energy profile from the variational theory (dashed), the single CSA (solid), and double CSA (long-dashed) are plotted as a function of N_f for three different persistence lengths in Fig. 10. The barrier heights from the simpler approximations are about twice the barriers obtained from the variational calculation for each persistence length. This is consistent with results of the other contiguous sequence approaches, where the barriers are much larger than obtained from simulations of $G\overline{o}$ models with pairwise additive forces.[9,10,62] In the present comparison, one could also expect this behavior, since a variational Hamiltonian with fewer degrees of freedom generally gives higher barriers (cf. the barriers in the two-dimensional illustration Fig. 5 with the full dimensional calculation in Fig. 7). Nevertheless, this comparison provides an intuitive explanation that agrees qualitatively with the magnitudes of the barrier heights. Confining the residues to be either folded or unfolded (in contiguous stretches) is responsible for the larger barrier heights, i.e., partial ordering reduces the barrier. This reduction seen in the more accurate calculation is reminiscent of the way wetting between two stable phases in nucleated phase transformations lowers nucleation barriers by reducing surface tension, and is discussed in the context of protein folding in Refs. 15 and 63.

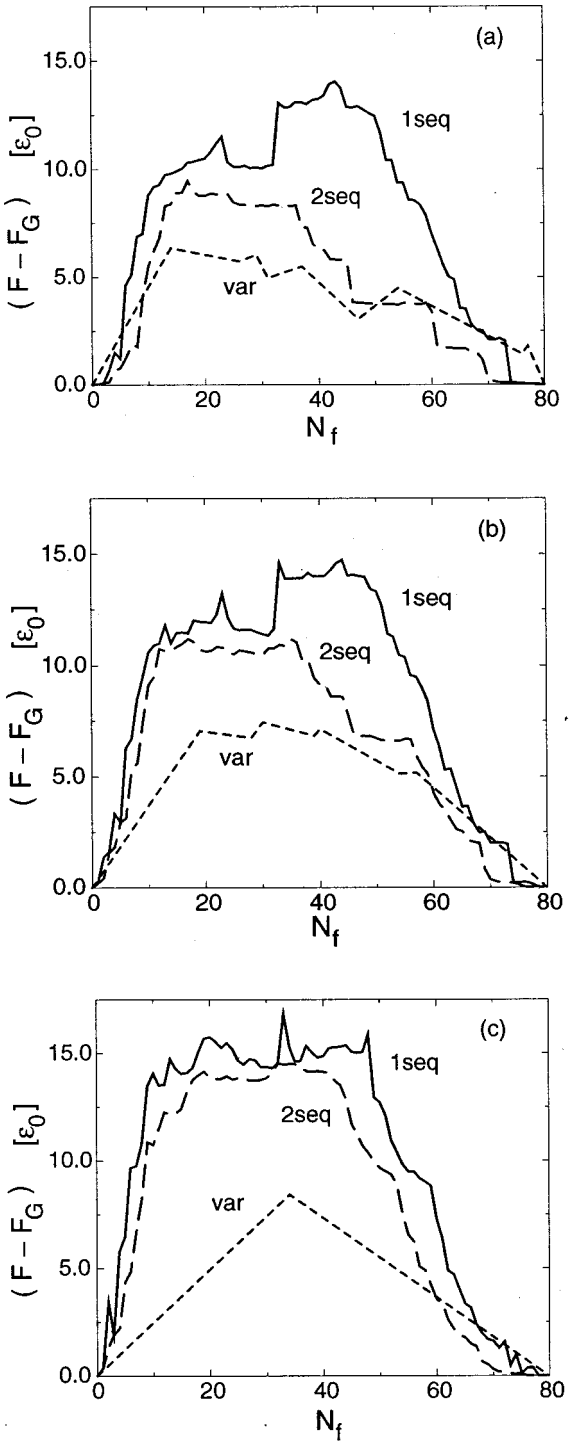

FIG. 10. Free energy vs number of folded residues for persistence lengths (a) $l=2a$ ($g=0.5$), (b) $l=5a$ ($g=0.8$), and (c) $l\approx 14a$ ($g=0.93$) scaled by the Miyazawa–Jernigan contact energy scale ϵ_0. The different curves correspond to contiguous sequence approximation (solid), double sequence approximation (long-dashed), and the variational theory (dashed). For the variational profile, N_f is defined by the number of residues with $\overline{\mathcal{B}_i}\geqslant 0.95$ (see text).

Considering the free energy profiles for the two approximations for a flexible chain ($l=2a$) plotted in Fig. 10(a), the maximum free energy for the single CSA is approximately 50% larger than the barrier from the double CSA and occurs at a greater value of N_f. Again there is no surprise in finding the barrier from the double contiguous approximation is

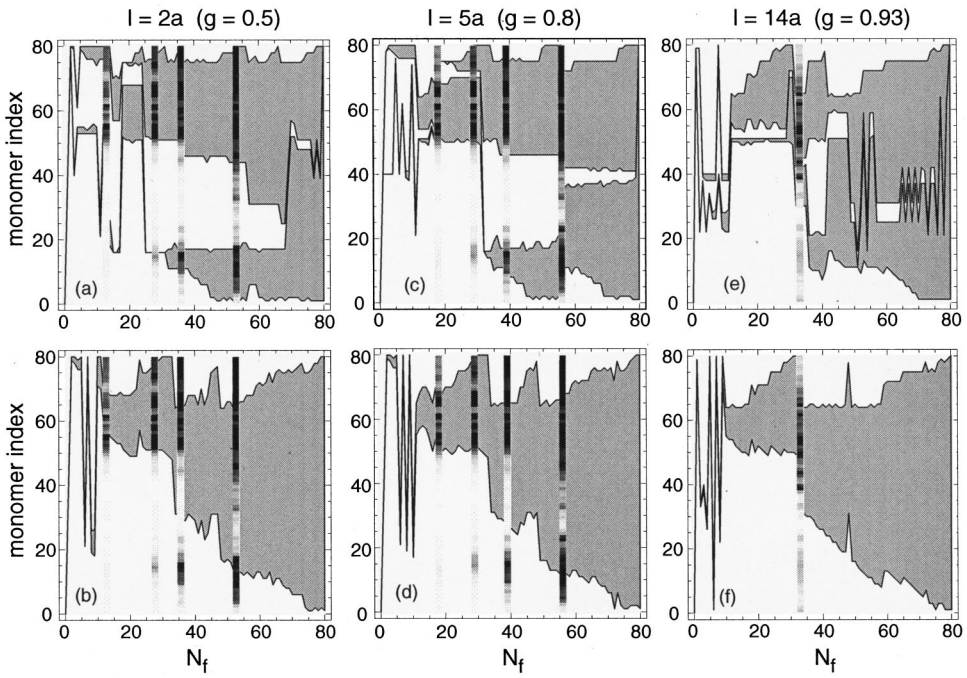

FIG. 11. Configuration of native residues vs number of folded resides persistence lengths for the double [top: (a), (c), (e)] and the contiguous [bottom: (b), (d), (f)] sequence approximations. The columns correspond to the persistence lengths: $l=2a$ (a), (b), $l=5a$ (c), (d), $l=14a$ (e), (f). Residues set to native constraints are indicated by the shaded region. The density plot corresponds to the normalized native density with $\alpha^N=0.5$ [Eq. (34)] evaluated at the transition states for each persistence length. (Black to white represents $\overline{\rho_i}=0$ to 1.) The ordinate for the density plot, N_f, is defined by the number of residues with $\overline{\mathcal{B}_i}\geqslant 0.95$ (see text).

lower, since the single contiguous configurations are a subset of the double contiguous configurations. It is interesting, however, to compare this difference as the chain stiffness increases. For a larger chain stiffness corresponding to polyalanine ($l=5a$) shown in Fig. 10(b), the difference in the barrier height from the single and double CSAs decreases, with the single CSA barrier approximately 30% larger than the double CSA barrier. Increasing the stiffness further ($l\approx 14a$), the free energy profiles shown in Fig. 10(c) are more similar, differing by only 10%. Since the profiles become still more similar as the chain stiffness increases (even though the conformations in the double CSA are less restricted), this comparison suggests that the single CSA more accurately describes stiff chain conformations than it does flexible chain conformations.

To make this connection more precise, we consider which residues are ordered along the folding route in these approximations. In the present treatment, this information can be easily represented by a plot of the folded regions specified by the monomer index as a function of N_f. The structured parts of the chain are indicated by the shaded regions in Fig. 11. To characterize the structure at the saddle points of the variational free energy surface we consider the Gaussian measure to the native structure

$$\rho_i=\left\langle \exp\left[-\frac{3}{2a^2}\alpha^N(\mathbf{r}_i-\mathbf{r}_i^N)^2\right]\right\rangle_0 =(1+\alpha^N G_{ii})^{-3/2}\exp\left[\frac{-3}{2a^2}\frac{\alpha^N(\mathbf{s}_i-\mathbf{r}_i^N)^2}{1+\alpha^N G_{ii}}\right]. \tag{33}$$

This measure of the monomer density relative to the native position is the order parameter employed in the companion paper to study the dynamics of the barrier crossing. The degree of native structure at the transition state can be characterized by the normalized measure

$$\overline{\rho_i}=(\rho_i-\rho_i^G)/(\rho_i^N-\rho_i^G), \tag{34}$$

where the superscripts G and N denote Eq. (33) evaluated at the constraints corresponding to the globule and native states, respectively.

Consider first the folded residues from the double CSA for a flexible chain ($l=2a$) shown in Fig. 11(a). Consistent with the full variational results, the shaded region clearly shows the structure forms between within helices H4–H5 and helix H1. Superimposed on the regions of folded residues (from the double CSA) are $\overline{\rho_i}$ as function of monomer index for the four main transition states of the folding route of a chain of the same stiffness (from the variational theory). The density plots of $\overline{\rho_i}$ indicate that the relatively unfolded regions are indeed more structured than the globule value $\overline{\rho_i}=0$. In particular, at the interface between the folded and unfolded regions in the double CSA, $\overline{\rho_i}$ obtains intermediate values indicating partial ordering. Nevertheless, the folded residues in the double CSA agrees qualitatively with the structure obtained from the variational saddle points. In contrast, the folded residues in the single CSA do not agree with the variational theory very well at all, as shown in Fig. 11(b).

Figures 11(c) and 11(d) show structured regions along the folding routes for a greater chain stiffness $l=5a$, the value appropriate for polyalanine. For this chain stiffness, the unfolded region between helix H1 and helices H4–H5 in the double CSA closes at a smaller value of N_f compared with the more flexible chain. In this sense the structured residues from the single and double CSAs are in better qualitative agreement, though the discrepancy between the two is still pronounced. The difference between the two approximations is greatly reduced for a chain with a much larger persistence length ($l\approx 14a$) shown in Figs. 11(e) and 11(f). While there are still unfolded regions between folded regions

in the double sequence approximation, they only persist for a very limited range of N_f. For this persistence length, the variational free energy surface has only one transition state as indicated by the density plot of ρ_i. The structure determined by the saddle point agrees qualitatively with the structure indicated by double CSA, but also with the single CSA (because the two are similar).

VI. CONCLUSION

In this paper, we used a variational approach to calculate the free energy profiles and characterize the transition state ensembles applicable when non-native contacts can be ignored. Using λ-repressor as an illustrative example, we investigated the role of chain stiffness on the fine structure of the free energy profile. We found that increasing the persistence length of the chain tends to smooth the free energy profile, making longitudinal ruggedness less pronounced. The transition state ensemble with very stiff chains was found to be more folded than the ensemble with more flexible chains. These results can be interpreted in terms of the tension between taking full advantage of strong local contacts while still respecting the bending rigidity of the chain. We also found that while the absolute barrier height has a pronounced maximum as a function of persistence length, the barrier scaled by the folding transition temperature $k_B T_f$ was relatively robust over a wide range of persistence lengths.

This study allowed us to investigate the applicability of simpler contiguous sequence approximations proposed recently. Both the free energy profiles and the configuration of folded residues along the folding routes suggest that the single CSA more accurately describes stiffer chains. Since the folded residues in the single CSA is roughly independent of the persistence length, the single CSA sequence approximation corresponds to a chain that has a longer effective persistence length than is appropriate for most natural proteins. The double CSA is able to capture the appropriate folded regions near the free energy barrier. On the other hand, the neglect of partial ordering leads to overestimates of the absolute barrier height for the λ-repressor protein. Nevertheless, these approximations should be good enough to compute the perturbations of the activated free energy found in protein engineering experiments that probe the transition state ensemble (ϕ analysis).[64]

ACKNOWLEDGMENT

This work has been supported by NIH Grant No. PHS 2 R01 GM44557.

APPENDIX

In this appendix, we outline a derivation of the expression for the entropy of the constrained chain given in Eq. (24).

The partition function of the harmonically constrained chain is given by

$$Z_0 = \exp\left[-A\sum_i C_i(\mathbf{r}_i^N)^2 \right] \times \mathrm{Tr}\exp\left[-A\sum_{ij} \mathbf{r}_i\cdot\Gamma^{(0)}\cdot\mathbf{r}_j + 2A\sum_i C_i\mathbf{r}_i^N\cdot\mathbf{r}_i \right], \tag{A1}$$

with $A=3/2a^2$ and Tr denotes $\int\Pi_i d\mathbf{r}_i$. These integrals are easy to evaluate by completing squares

$$Z_0 = (\pi/A)^{3n/2}(\det G)^{3/2} \times \exp\left[-A\sum_i C_i(\mathbf{r}_i^N)^2 + A\sum_{ij} C_i\mathbf{r}_i^N\cdot G_{ij}\cdot C_j\mathbf{r}_j^N \right], \tag{A2}$$

where $G=\Gamma^{(0)-1}$. We note that Eq. (15) can be used to express the second term in the exponent as

$$A\sum_{ij} C_i\mathbf{r}_i^N\cdot G_{ij}C_j\mathbf{r}_j^N = A\sum_i C_i\mathbf{r}_i^N\cdot\mathbf{s}_i \tag{A3}$$

in terms of the average monomer positions $\{\mathbf{s}_i\}$. From Eq. (20), combining the partition function with

$$A\sum_i C_i\langle(\mathbf{r}_i-\mathbf{r}_i^N)^2\rangle_0 = A\sum_i C_i(G_{ii}a^2+\mathbf{s}_i^2 + 2\mathbf{s}_i\cdot\mathbf{r}_i^N+(\mathbf{r}_i^N)^2) \tag{A4}$$

gives the entropy of the constrained chain (ignoring a constant factor)

$$S[\{C\}] = \frac{3}{2}\log\det G + \frac{3}{2}\sum_i C_iG_{ii} + A\sum_i C_i\mathbf{s}_i^2 - A\sum_i C_i\mathbf{r}_i^N\cdot\mathbf{s}_i. \tag{A5}$$

This expression is easier to interpret after simplifying the last two terms. Inserting the identity, $G\Gamma^{(0)}=\mathbf{1}$, into the last term and using Eq. (15) gives

$$A\sum_i C_i\mathbf{r}_i^N\cdot\mathbf{s}_i = A\sum_{ikj} C_i\mathbf{r}_i\cdot G_{ik}\Gamma_{kj}^{(0)}\cdot\mathbf{s}_j = A\sum_{kj}\mathbf{s}_k\cdot\Gamma_{kj}^{(0)}\cdot\mathbf{s}_j. \tag{A6}$$

In this form, the last two terms of Eq. (A5) can be combined

$$-A\sum_{kj}\mathbf{s}_k\cdot(\Gamma_{kj}^{(0)}-C_k\delta_{kj})\cdot\mathbf{s}_j = -A\sum_{kj}\mathbf{s}_k\cdot\Gamma_{kj}^{(\mathrm{ch})}\cdot\mathbf{s}_j, \tag{A7}$$

where we have used the definition of $\Gamma^{(0)}$ given in Eq. (17). Thus, we have

$$S[\{C\}] = \frac{3}{2}\log\det G + \frac{3}{2}\sum C_iG_{ii} - \frac{3}{2a^2}\sum_{kj}\mathbf{s}_k\cdot\Gamma_{kj}^{(\mathrm{ch})}\cdot\mathbf{s}_j, \tag{A8}$$

which is Eq. (24).

[1] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).
[2] J. D. Bryngelson and P. G. Wolynes, J. Phys. Chem. **93**, 6902 (1989).
[3] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins: Struct., Funct., Genet. **21**, 167 (1995).
[4] D. Thirumalai, V. Ashwin, and J. K. Bhattacharjee, Phys. Rev. Lett. **77**, 5385 (1996).
[5] E. G. Timoshenko, Y. A. Kuznetsov, and K. A. Dawson, Phys. Rev. E **54**, 4071 (1996).
[6] S. Takada, J. J. Portman, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **94**, 2318 (1997).
[7] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721 (1992).
[8] J. N. Onuchic, P. G. Wolynes, N. D. Socci, and Z. A. Luthey-Schulten, Proc. Natl. Acad. Sci. U.S.A. **92**, 3626 (1995).
[9] H. Nymeyer, A. E. Garcia, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **95**, 5921 (1998).
[10] J.-E. Shea, J. N. Onuchic, and C. L. Brooks III, Proc. Natl. Acad. Sci. U.S.A. **96**, 12512 (1999).
[11] C. Hardin, Z. Luthey-Schulten, and P. G. Wolynes, Proteins: Struct., Funct., Genet. **34**, 281 (1999).
[12] V. S. Pande and D. S. Rokhsar, Proc. Natl. Acad. Sci. U.S.A. **96**, 1273 (1999).
[13] M. Sasai and P. G. Wolynes, Phys. Rev. A **46**, 7979 (1992).
[14] S. Takada and P. G. Wolynes, Phys. Rev. E **55**, 4562 (1997).
[15] S. Takada and P. G. Wolynes, J. Chem. Phys. **107**, 9585 (1997).
[16] J. J. Portman, S. Takada, and P. G. Wolynes, Phys. Rev. Lett. **81**, 5237 (1998).
[17] J. J. Portman, S. Takada, and P. G. Wolynes, J. Chem. Phys. **114**, 5082 (2001), following paper.
[18] P. L. Privalov, Adv. Protein Chem. **33**, 167 (1979).
[19] H. Frauenfelder, G. A. Petsko, and D. Tsernoglou, Nature (London) **280**, 558 (1979).
[20] D. W. Oxtoby, in *Liquids, Freezing and the Glass Transition*, edited by D. Levesque, J. P. Hansen, and J. Zinn-Justin (Elsevier, New York, 1991), p. 149–191.
[21] R. E. Burton, G. S. Huang, M. A. Daugherty, P. W. Fullbright, and T. G. Oas, J. Mol. Biol. **263**, 311 (1996).
[22] R. E. Burton, G. S. Huang, M. A. Daugherty, T. L. Calderone, and T. G. Oas, Nat. Struct. Biol. **4**, 305 (1997).
[23] P. J. Flory, *Statistical Mechanics of Chain Molecules* (Wiley, New York, 1969).
[24] M. Bixon and R. Zwanzig, J. Chem. Phys. **68**, 1896 (1978).
[25] D. Thirumalai and B. Y. Ha, in *Theoretical and Mathematical Models in Polymer Research*, edited by A. Y. Grossberg (Academic, Boston, 1998), pp. 1–35.
[26] A. Perico, Biopolymers **28**, 1527 (1989).
[27] A. Perico, F. Ganazzoli, and G. Allegra, J. Chem. Phys. **87**, 3677 (1987).
[28] A. Perico, Acc. Chem. Res. **22**, 336 (1989).
[29] Y. Hu, G. R. Fleming, K. F. Freed, and A. Perico, Chem. Phys. **158**, 395 (1991).
[30] A. Perico, N. E. Moe, and M. D. Ediger, J. Chem. Phys. **108**, 1245 (1998).
[31] H. Yamakawa, *Helical Wormlike Chains in Polymer Solutions* (Springer-Verlag, Berlin, 1997).
[32] C. R. Cantor and P. R. Schimmel, *Biophysical Chemistry* (W. H. Freeman, New York, 1980), Vol. 3.
[33] P. J. Flory, Proc. R. Soc. London A **351**, 351 (1976).
[34] M. P. Solf and T. A. Vilgis, J. Phys. A **28**, 6655 (1995).
[35] J. D. Bryngelson and D. Thirumalai, Phys. Rev. Lett. **76**, 542 (1996).
[36] M. P. Solf and T. A. Vilgis, Phys. Rev. Lett. **77**, 4276 (1996).
[37] R. Zwanzig, J. Chem. Phys. **106**, 2824 (1996).
[38] T. A. Vilgis, Macromol. Theory Simul. **7**, 59 (1998).
[39] M. P. Solf and T. A. Vilgis, Phys. Rev. E **55**, 3037 (1997).
[40] I. Bahar, A. R. Atilgan, and B. Erman, Folding Des. **2**, 173 (1997).
[41] I. Bahar, A. R. Atilgan, M. C. Demirel, and B. Erman, Phys. Rev. Lett. **80**, 2733 (1998).
[42] I. Bahar, A. Wallqvist, and D. B. Covell, Biochemistry **37**, 1067 (1998).
[43] B. Erman and K. A. Dill, J. Chem. Phys. **112**, 1050 (2000).
[44] B. A. Shoemaker, J. Wang, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **94**, 777 (1997).
[45] B. A. Shoemaker, J. Wang, and P. G. Wolynes, J. Mol. Biol. **287**, 675 (1999).
[46] O. V. Galzitskaya and A. V. Finkelstein, Proc. Natl. Acad. Sci. U.S.A. **96**, 112999 (1999).
[47] E. Alm and D. Baker, Proc. Natl. Acad. Sci. U.S.A. **96**, 11305 (1999).
[48] V. Muñoz and W. A. Eaton, Proc. Natl. Acad. Sci. U.S.A. **96**, 11311 (1999).
[49] H. S. Chan and K. A. Dill, J. Chem. Phys. **90**, 492 (1990).
[50] H. S. Chan and K. A. Dill, J. Chem. Phys. **92**, 3118 (1990).
[51] R. G. Parr and W. Yang, *Density Functional Theory of Atoms and Molecules* (Oxford University Press, Oxford, 1989).
[52] D. J. Wales, J. Chem. Phys. **101**, 3750 (1994).
[53] Throughout this section, the sequence index begins with the first residue of the fast folding double mutant (G46A/G48A) of λ_{6-85} (Ref. 21). The helices roughly correspond to the sequence indices H1(3–21), H2(27–36), H3(38–47), H4(53–65), H5(72–80) as determined by the DSSP algorithm (Ref. 55) applied to the PDB structure (Ref. 54).
[54] L. J. Beamer and C. O. Pabo, J. Mol. Biol. **227**, 177 (1992).
[55] W. Kabsch and C. Sander, Biopolymers **22**, 2577 (1983).
[56] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1996).
[57] As rough estimate of R_G with no confinement $B=0$, the asymptotic expression (Ref. 23) $R_G^2/(n-1)a^2 \sim (1/6)(1+g)/(1-g)$ gives $R_G \approx 20a$ for a free chain of length $n=80$ and persistence length $l=20a$ ($g=0.95$).
[58] The reported values for the persistence length are determined through characteristic ratio defined as $C_\infty = \lim_{n\to\infty}\langle(\mathbf{r}_n-\mathbf{r}_1)^2\rangle/na^2$ (see Refs. 23 and 32). For an infinite freely rotating chain with chain stiffness g, $C_\infty = (1+g)/(1-g)$; using $l=a/(1-g)$ gives $l=(C_\infty+1)a/2$.
[59] M. Silow and M. Oliveberg, J. Mol. Biol. **269**, 611 (1997).
[60] S. S. Plotkin, J. Wang, and P. G. Wolynes, Phys. Rev. E **53**, 6271 (1996).
[61] The configurations in the contiguous sequence approximations are specified by the constraint parameters $\{C_i\}$: if the ith residue is fully folded (unfolded), C_i is set to the value of the corresponding constraint at the native (globule) minimum.
[62] M. P. Eastwood and P. G. Wolynes, J. Chem. Phys. (in press).
[63] P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **94**, 6170 (1997).
[64] A. R. Fersht, A. Matouschek, and L. Serrano, J. Mol. Biol. **224**, 771 (1992).