# Variational Theory for Site Resolved Protein Folding Free Energy Surfaces

J. J. Portman, S. Takada,* and P. G. Wolynes

*Departments of Physics and Chemistry, University of Illinois, Urbana, Illinois, 61801*

(Received 28 May 1998)

We present a microscopic variational theory for the free energy surface of a fast folding protein that allows folding kinetics to be resolved to the residue level using Debye-Waller factors as local order parameters. We apply the method to the $\lambda$-repressor protein and compare with site directed mutagenesis experiments. The formation of native structure and the free energy profile along the folding route are shown to be well described by the capillarity approximation but with some fine structure due to local folding topology. [S0031-9007(98)07855-7]

PACS numbers: 87.15.–v

Proteins fold on a configurational energy landscape that has the shape of a funnel [1]. As the protein moves down the funnel towards the native state, incomplete cancellation of the entropy and energy losses may result in free energy barriers. So far, proteins that fold fast exhibit single exponential kinetics [2], consistent with a free energy profile that has a single highest barrier along the progress coordinate. Central issues are the origin of the free energy barrier for fast folding proteins and how the ensemble of structures which represent the bottleneck is to be characterized. We address these questions using a variational approximation that describes ensembles of partially folded proteins at the highest level of resolution, i.e., the specific role of individual residues in guiding the protein to the native state is quantified. In the laboratory, Fersht has developed a probe of the transition state or bottleneck ensemble through protein engineering kinetic studies in which the sequence of the protein is altered by replacing residues one at a time [3]. The experiment yields the fraction of the time that the mutated site is in the native conformation in the bottleneck ensemble by comparing folding rates of the mutant to the wild type. Since this can be done for any residue in the sequence, these studies are inherently “site resolved.” Resolving the transition state ensemble to this level is one way to monitor the average of the many routes taken as proteins fold.

Previous analytic mean field theories and simulations have produced energy landscapes in one or two global dimensions characterizing the folding ensemble [4]. We develop here a free energy profile for proteins with a funneled landscape that is completely site resolved, i.e., one dimension per residue, by extending the mean field variational calculations presented in [5]. The underlying Hamiltonian explicitly incorporates chain stiffness and connectivity while the approximation employs a variational density that monitors local order parameters for folding akin to the Debye-Waller factors (also called temperature factors) for individual residues seen in x-ray crystallography.

The basic Hamiltonian for an interacting polymer chain is $H = H_{\mathrm{chain}} + H_{\mathrm{int}}$, where $H_{\mathrm{chain}}$ is backbone potential and $H_{\mathrm{int}}$ are the interactions between distant monomers along the chain. $H_{\mathrm{chain}}$ is an effective harmonic potential $\beta H_{\mathrm{chain}} = 1/2 \sum \mathbf{r}_i \cdot \Gamma_{ij} \cdot \mathbf{r}_j + B \sum \mathbf{r}_i^2$, where $\{\mathbf{r}_i\}$ are the positions of the $N$ $\alpha$ carbons, and $\beta = 1/k_B T$ is the inverse temperature. The first term enforces the chain connectivity while the second term confines the radius of gyration to a reasonable value [achieved by fixing $B$ to a small constant: $B = (3/2a^2) \times 10^{-3}$]. For the connectivity matrix, $\Gamma_{ij}$, we use the well known Gaussian approximation to the freely rotating chain derived in [6]. Denoting the $i$th bond vector by $\mathbf{a}_i = (\mathbf{r}_{i+1} - \mathbf{r}_i)$ and the angle between successive bond vector by $\theta$, this stiff chain model is defined by the correlations $\langle \mathbf{a}_i \cdot \mathbf{a}_{i+l} \rangle = a^2 g^l$, where $a$ is the mean bond length and $g = \cos\theta$. Following Bixon and Zwanzig, $\Gamma_{ij}$ is determined by inverting these correlations and transforming to the bead representation resulting in a pentadiagonal matrix that depends on the stiffness parameter $g$ (for the explicit matrix, see [6]). In the limit $g \to 0$, $\Gamma_{ij}$ describes the standard flexible chain, whereas $g \to 1$ corresponds to a rigid rod. The persistence length, $l$, for this chain is given by $l = a/(1 - g)$. We use $g = 0.8$ giving (with $a = 3.8$ Å) $l \approx 20$ Å, the persistence length for poly-$L$-alanine [7].

We take interaction between distant monomers along the chain to be restricted to specifically nativelike interactions $H_{\mathrm{int}} = \sum^{(N)} \epsilon_{ij} u(|\mathbf{r}_i - \mathbf{r}_j|)$, where the isotropic pair potential has a minimum at a nonzero distance produced by summing three Gaussians, $u(r) = \gamma_s e^{-\beta_s r^2} + \gamma_i e^{-\beta_i r^2} - \gamma_l e^{-\beta_l r^2}$; the short- and intermediate-range terms are repulsive while the long-range Gaussian is attractive ($\beta_s > \beta_i > \beta_l$). The sum over pair interactions ($\sum^{(N)}$) is restricted to native contacts. This constraint gives a smooth funnel shaped energy landscape, appropriate for fast folding proteins. This is an extreme realization of the *principle of minimum frustration* [8] and is reminiscent of the lattice model originally introduced by Gō [9]. The heterogeneity of the interaction between different residues is reflected by the strength $\epsilon_{ij}$. Non-native interactions can also be included in $H_{\mathrm{int}}$ and treated by our variational method.

To study the many dimensional free energy surface defined by $H$, we choose local order parameters that can characterize the ensemble of partially folded structures by specifying the temperature factor for each residue, $\hat{\mathcal{B}}_i$. This describes the mean square fluctuations of a residue about its native position and for fully folded proteins has been measured. A similar local order parameter for folding has been used in lattice simulations [10] and earlier analytical work [5]. Consider the free energy surface defined by the set of scalar fluctuations of each residue from its native position $\{\mathbf{r}_i^N\}$, $\hat{\mathcal{B}}_i = (\mathbf{r}_i - \mathbf{r}_i^N)^2$. The free energy surface $\mathcal{F}[\{\hat{\mathcal{B}}_i\}]$ for an ensemble specified by $\{\hat{\mathcal{B}}_i\}$ is given by

$$e^{-\beta \mathcal{F}[\{\hat{\mathcal{B}}_i\}]} = \mathrm{Tr}\left[\prod_i \delta[\hat{\mathcal{B}}_i - (\mathbf{r}_i - \mathbf{r}_i^N)^2] e^{-\beta H}\right] \quad (1)$$

$$= \int \mathcal{D}\lambda \, \mathrm{Tr}\, e^{-\beta H[\lambda]}, \quad (2)$$

where $\beta H[\lambda] = \beta H + \sum \lambda_i[\hat{\mathcal{B}}_i - (\mathbf{r}_i - \mathbf{r}_i^N)^2]$, and $\mathcal{D}\lambda \equiv \prod_j d\lambda_j/2\pi i$.

Denoting the integrand in Eq. (2) by $e^{-\beta F[\lambda]}$, we approximate $F[\lambda]$ with the help of a reference Hamiltonian $H_0$ and the Gibbs-Bogoliubov variational expression $F[\lambda] \approx -k_B T \log Z_0 + \langle H[\lambda] - H_0 \rangle_0$, where $Z_0 = \mathrm{Tr}[e^{-\beta H_0}]$, and $\langle \cdots \rangle_0$ means the average with respect to $H_0$. The reference Hamiltonian describes a Gaussian chain constrained to fluctuate about the native structure $\{\mathbf{r}_i^N\}$ by a harmonic external field: $\beta H_0 = H_{\mathrm{chain}} + \sum C_i(\mathbf{r}_i - \mathbf{r}_i^N)^2$. The variational parameters $\{C_i\}$ are conjugate to $\{\hat{\mathcal{B}}_i\}$. This reference Hamiltonian captures the two stable phases of fast folding proteins: the globule with small $\{C_i\}$ and the native state with uniformly large $\{C_i\}$. $\{C_i\}$ also form a set of local order parameters for folding. A similar (but more elaborate) reference Hamiltonian was used to determine the phase diagram for proteins with a rugged energy landscape as well as to study folding free energy barriers [5]. These mean field studies employed a global order parameter for nativeness by setting all $C_i$ equal in one region of the protein. Different related effective harmonic variational Hamilitonians have been employed to study polymers in random media [11], random directed polymers [12], and random copolymers [13].

Evaluating Eq. (2) using the steepest descents approximation gives the stationary condition $\hat{\mathcal{B}}_i = \langle(\mathbf{r}_i - \mathbf{r}_i^N)^2\rangle_0$ as a function of $\{C_i\}$ leading to a one to one relation between the $\{\hat{\mathcal{B}}_i\}$ and $\{C_i\}$. Technically, it is more convenient to study the free energy surface in $\{C_i\}$ space so that we consider the variational free energy surface expressed as $F[\{C_i\}] = E - ST$ with the estimate for the energy $E = \sum^{(N)} \epsilon_{ij} \langle u(\mathbf{r}_{ij})\rangle_0$ and the entropy $S/k_B = \log Z_0 + \sum \langle C_i(\mathbf{r}_i - \mathbf{r}_i^N)^2\rangle_0$ as a function of $\{C_i\}$.

Since $H_0$ is quadratic, $Z_0$ and all of the averages are expressible in terms of the correlations, $\langle \mathbf{r}_i \cdot \mathbf{r}_j\rangle_0 - \langle \mathbf{r}_i\rangle_0 \cdot \langle \mathbf{r}_j\rangle_0 = 3/2G_{ij}$, with $G_{ij} = [1/2\ \Gamma_{ij} + (B + C_i)\delta_{ij}]^{-1}$. $Z_0$ involves the determinant of the correlation matrix while the averages can be calculated using the density of site $i$, $\rho_i(\mathbf{r}) = \langle \delta(\mathbf{r} - \mathbf{r}_i)\rangle_0$, and the pair density between sites $i$ and $j$, $\rho_{ij}(\mathbf{r}) = \langle \delta(\mathbf{r} - \mathbf{r}_{ij})\rangle_0$. In terms of the average position $\mathbf{s}_i = \sum_j G_{ij} C_j \mathbf{r}_j^N$, these densities are $\rho_i(\mathbf{r}) = (\pi G_{ii})^{-3/2} \exp[-(\mathbf{r} - \mathbf{s}_i)^2/G_{ii}]$, and $\rho_{ij}(\mathbf{r}) = (\pi \delta G_{ij})^{-3/2} \exp[-(\mathbf{r} - \mathbf{s}_{ij})^2/\delta G_{ij}]$, where $\delta G_{ij} = G_{ii} + G_{jj} - 2G_{ij}$.

Since both $F[\{C_i\}]$ and $\nabla_C F[\{C_i\}]$ can be expressed analytically in terms of $G_{ij}$ (which is calculated numerically), it is relatively easy to locate the minima and saddle points numerically [14]. Once the transition states and the folded, unfolded, and local minima are determined, we define the average folding route to be the connected steepest descents path from each transition state to the neighboring minima. Only the global minimum of $F[\{C_i\}]$ is rigorously an upper bound, but the saddle points and local minima should also be good estimates for the true free energy surface.

We now apply the model to the folding of the $\lambda$-repressor protein. $\lambda_{6-85}$ is a good candidate system since it is small (80 residues) and folds extremely rapidly in 20 $\mu$s following two-state kinetics [15]. Recently Oas *et al.* probed the structure of the transition state ensemble of $\lambda_{6-85}$ by comparing the folding rates measured with NMR for seven mutants made by alanine to glycine replacements [16]. The folding rates can be connected to the structure of the transition state by the $\phi$ parameter developed by Fersht [3], $\phi = \Delta \log k_f / \Delta \log K$ ($k_f$ is the folding rate and $K$ denotes the equilibrium constant). $\phi \sim 0$ indicates that the conformation of the mutated residue in the transition state is similar to the globule, whereas $\phi \sim 1$ suggests that this residue has native structure in the transition state ensemble. Based in part on these $\phi$ values, Table I, Oas proposed that helices H1 and H4 are structured in the transition state ensemble. While a more extensive mutation study is necessary to characterize fully the transition state ensemble, a comparison to these results is a strong test for the theory presented here.

We define native contacts between residues with $\beta$ carbons within a distance of 6.5 Å ($\alpha$ carbons for glycines) in the native structure [17] that are separated by at least four monomers in sequence. The pair distribution function of the distance between $\alpha$ carbons is used to constrain the parameters of the effective pair potential. We find the intermediate- and long-ranged interaction parameters $(\gamma_i, \beta_i a^2, \gamma_l, \beta_l a^2) = (9.0, 0.8, 6.0, 0.4)$ give an effective potential well that contains all the native contact distances and has a minimum at the most probable $\mathrm{C}_\alpha - \mathrm{C}_\alpha$ contact distance, $r^* = 1.6a$, with $u(r^*) = -1$.

TABLE I. $\phi$ values for the $\lambda$-repressor protein.

| Mutant<br>(Helix) | M15<br>(H1) | M20<br>(H1) | M37<br>(H2) | M49<br>(H3) | M63<br>(H4) | M66<br>(H4) | M81<br>(H5) |
|---|---|---|---|---|---|---|---|
| $\phi_{\mathrm{Exp}}$ | 0.5 | 1.0 | 0.2 | 0.3 | 0.8 | 1.2 | 0.6 |
| $\langle\phi\rangle_{\mathrm{Calc}}$ | 0.3 | 0.3 | 0.1 | 0.2 | 1.0 | 1.0 | 0.7 |

The short-range interaction represents the hard-core repulsion between residues and gives excluded volume; with the choice of values $(\gamma_s, \beta_s a^2) = (25.0, 4.5)$, the repulsion roughly balances the attractive energy in the globule state allowing us to study a folding transition that occurs directly from a random coil (i.e., near the theta temperature). We will compare the free energy surface using a homogeneous contact strength ($\epsilon_{ij} = \epsilon_0$) with that of the full 20 letter Miyazawa-Jernigan contact energies [18] in which contact between different residues have different energies.

We now consider a low energy folding route on the free energy surface connecting the globule and native minima. Figure 1 shows the free energy along this path at the folding transition temperature, $T_f$, plotted as a function of the fraction of energy stabilization relative to the native state, $E_{\mathrm{NORM}}$. The stationary points of the free energy surface form a broad barrier with a reasonable height $[(5-7)k_B T_f]$ for fast folding proteins. The barrier for the homogeneous case of all equal interactions is approximately 30% larger than for the heterogeneous contact energy model. Another difference between the two models is that we find many more transition states and local minima (not shown) for the heterogeneous case. These arise from the competition between contacts of different strengths.

The Debye-Waller factor (temperature factor) of each residue contains structural information of the stationary points along the folding route. The temperature factors plotted versus sequence number at four of the saddle points along the folding route for the heterogeneous model are shown in Fig. 2. Even the globule already has some structure though its fluctuations are large. Comparing these curves progressively from the globule to the native state, we see that the barrier at $TS_1$ is described by the formation of helices H4 and H5 while the central region of helix H1 which docks with H4 is partially localized but with substantial fluctuations. This suggests that the stabilizing contacts between H4, H5, and H1 are due to the general increase in density rather than any very strong contacts between specific residues. Following this is the completion of helix H5 and the center of helix H1, while helices H2 and H3 remain relatively disordered at $TS_3$. Last, helices H2 and H3 become increasingly more ordered along this folding route as indicated by the temperature factors at $TS_4$.

The folding route described above agrees with the conclusions of the Oas group [16]; namely, helices H1 and H4 are structured in the transition state ensemble, whereas helices H2 and H3 are unstructured. The $\phi$ values obtained from this calculation makes the comparison more precise. As in the experimental analysis, we assume the ensemble of structures does not change but recalculate the free energy for each mutant at the saddle points. Using $k \sim e^{-\beta \Delta F^{\dagger}}$, we calculate $\phi$ at each saddle point and their average over the four transition states. The results are given in Table I. The agreement with experiment is quite reasonable in light of the rough approximations made in modeling the experiment. The worst agreement is for the mutation M20. This is a surface residue with no tertiary contacts by our definition; thus other terms in the energy may be contributing. Some obvious improvements to this model such as explicit hydrogen bonding and many body forces can easily be made, but our aim here is to explore the simplest model that gives a physically reasonable and direct picture of the folding route for fast folding proteins. From this point of view, the agreement with experiment is very encouraging.

Examination of the average folding route also leads to a simple physical picture for the barriers under the

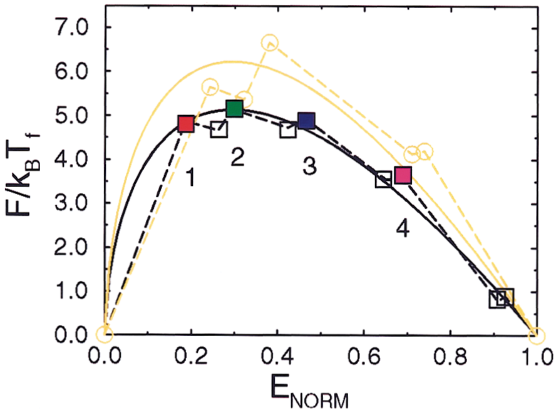

FIG. 1(color). The free energy at the stationary points along the folding route as a function of the normalized energy for the homogeneous (orange ○) and inhomogeneous (black □) models. For an ensemble with average energy $E$, $E_{\mathrm{NORM}} = (E - E_G)/(E_N - E_G)$, where $E_N$ and $E_G$ are the energies of the native state and globule state, respectively. $f_{\mathrm{cap}}$ (described in the text) is also shown as the solid line with $\gamma = 35(42)k_B T_f$ for the heterogeneous (homogeneous) model.

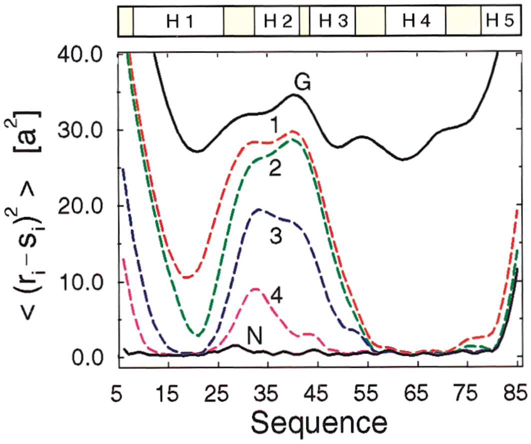

FIG. 2(color). The temperature factors (i.e., mean square fluctuations relative to the average position of $i$th monomer, $\mathbf{s}_i$) plotted as a function of sequence number for the heterogeneous model at the stationary points shown in Fig. 1. The bar at the top indicates the helical secondary structure (H1–H5).

thermodynamic conditions of folding at $T_f$ directly from the random coil. The progression of the folding of the 3D structure is shown in Fig. 3, where the sites of the native structure are colored according to the fraction of energy gained at that site. The first bottleneck involves partial structure formation in approximately 40% of the chain (in helices H4 and H5). Subsequently, a picture much like that of the growth of an ordered phase in an ordinary first order transition emerges with a front of progressive ordering crossing the protein. This is reminiscent of the capillarity theory [19]. Within the capillarity picture, one imagines an ordered region that is completely folded separated by a sharp interface from a completely unfolded region. At $T_f$, the free energy of progressively forming folded structure is given by $f_{\mathrm{cap}} = \gamma(-N_f + N_f^{2/3})$, where $N_f$ is the fraction of native residues, and $\gamma$ is the surface energy cost. As shown in Fig. 1, this equation provides a good fit to the stationary points in both the homogeneous and heterogeneous models, identifying $N_f$ with normalized energy $E_{\mathrm{Norm}}$ (defined in Fig. 1) and treating $\gamma$ as a fitting parameter.

Superimposed on the average behavior of the profile are fluctuations representing the fine structure arising from inhomogeneity of the local folding free energy. It is obvious that these fluctuations arise for the heterogeneous model because of varying interaction energies but are still present for the pure homogeneous Gō-like model. This shows the high free energy intermediates [10,20] along the average folding route for a very funnellike surface are mostly determined by the folded topology. Within the capillarity picture, the smaller barrier for the heterogeneous case can be interpreted as being due to wetting, as expected for the random field Ising model [21]. The thermodynamic conditions studied here favor the capillarity picture with a sharp interface. When folding occurs from an already collapsed state the free energy difference of the bulk unfolded and folded phases is smaller leading to a broader interface. The basic formalism can be used for this other regime as well.

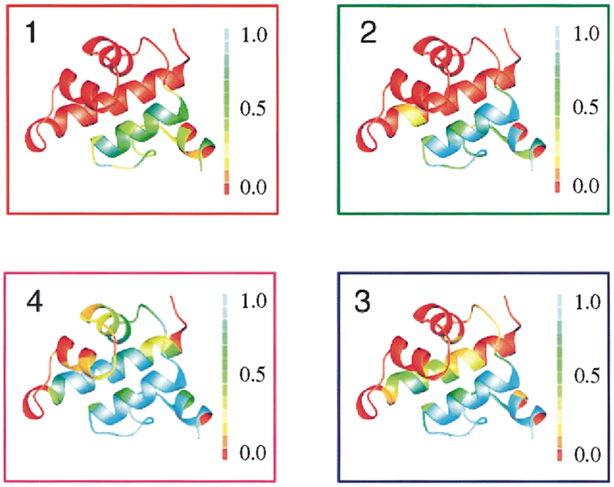

FIG. 3(color). The 3D native structure of the $\lambda$-repressor protein colored according to the normalized energy of each site, $(e_i - e_i^G)/(e_i^N - e_i^G)$ with $e_i = \sum_j^{(N)} \epsilon_{ij}\langle u(r_{ij})\rangle_0$, evaluated at each saddle point (clockwise from upper left). $e_i^N$ and $e_i^G$ are the energies of the $i$th site in the native state and globule state, respectively.

We thank Ben Shoemaker and Michael Eastwood for helpful discussions. This work was supported by NIH Grant No. PHS R01 GM44557. S. T. was supported by the Japan Society for Promotion of Science.

*Present address: Department of Chemistry, Kobe University, Rokkodai, Kobe, 657 Japan.

[1] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721 (1992); J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins Struct. Funct. Genet. **21**, 167 (1995).

[2] W. A. Eaton *et al.*, Curr. Opin. Struct. Biol. **7**, 10 (1997).

[3] A. R. Fersht, A. Matouschek, and L. Serrano, J. Mol. Biol. **224**, 771 (1992).

[4] S. S. Plotkin, J. Wang, and P. G. Wolynes, J. Chem. Phys. **106**, 2932 (1997); J. N. Onuchic, N. D. Socci, Z. A. Luthey-Schulten, and P. G. Wolynes, Folding Design **1**, 441 (1996); Z. Guo, C. L. Brooks, and E. M. Boczko, Proc. Natl. Acad. Sci. U.S.A. **94**, 10 161 (1997).

[5] M. Sasai and P. G. Wolynes, Phys. Rev. A **46**, 7979 (1992); S. Takada and P. G. Wolynes, Phys. Rev. E **55**, 4562 (1997); S. Takada and P. G. Wolynes, J. Chem. Phys. **107**, 9585 (1997).

[6] M. Bixon and R. Zwanzig, J. Chem. Phys. **68**, 1896 (1978).

[7] C. R. Cantor and P. R. Schimmel, *Biophysical Chemistry* (W. H. Freeman and Co., New York, 1980), Vol. 3, p. 1013.

[8] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).

[9] N. Gō, Annu. Rev. Biophys. Bioeng. **12**, 183 (1983).

[10] V. Pande and D. S. Rokhsar (to be published).

[11] S. F. Edwards and M. Muthukumar, J. Chem. Phys. **89**, 2435 (1988); J. D. Honeycutt and D. Thirumalai, J. Chem. Phys. **90**, 4542 (1989).

[12] M. Mézard and G. Parisi, J. Phys. I (France) **1**, 809 (1991); T. Garel and H. Orland, Phys. Rev. B **55**, 226 (1997).

[13] A. Moskalenko, Y. A. Kuznetsov, and K. A. Dawson, Physica (Amsterdam) **249A**, 353 (1998).

[14] D. J. Wales, J. Chem. Phys. **101**, 3750 (1994).

[15] R. E. Burton *et al.*, J. Mol. Biol. **263**, 311 (1996).

[16] R. E. Burton *et al.*, Nature Struct. Biol. **4**, 305 (1997).

[17] L. J. Beamer and C. O. Pabo, J. Mol. Biol. **227**, 177 (1992).

[18] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1996).

[19] J. D. Bryngleson and P. G. Wolynes, Biopolymers **30**, 177 (1990); A. V. Finkelstein and A. Y. Badretdinov, Folding Design **2**, 115 (1997); P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **94**, 6170 (1997).

[20] M.Silow and M. Oliveberg, J. Mol. Biol. **269**, 611 (1997).

[21] J. Villain, J. Phys. I (France) **46**, 1843 (1985).