Compaction and folding in model proteins

Ting-Lan Chiu
Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109-1055

Richard A. Goldstein[a)]
Department of Chemistry and Biophysics Research Division, University of Michigan, Ann Arbor, Michigan 48109-1055

(Received 9 December 1996; accepted 11 June 1997)

Protein folding is modeled as diffusion on a free-energy landscape, allowing use of the diffusion equation to study the impact of energetic parameters on the folding dynamics. The free-energy landscape is characterized by two different order parameters, one representing the degree of compactness, the other a measure of the progress towards the folded state. For marginally stable proteins, fastest folding is achieved when the nonspecific interactions favoring compaction are strong, resulting in a high folding temperature. Such proteins fold by rapid collapse followed by slower accumulation of correct contacts. © *1997 American Institute of Physics.* [S0021-9606(97)50435-1]

I. INTRODUCTION

The mechanism by which proteins fold into their native three-dimensional structures remains one of the central unsolved problems of molecular biophysical chemistry. Structural organization occurs on a variety of size scales, including the generation of local structure with specific packing interactions, the formation of the correct topology, and an overall compaction. A number of phenomenological models have been proposed that attempt to understand the relationship between these various processes that comprise folding. These models range from those proposing that local structures form first, followed by the assembly of these preformed elements, as in the ''framework model'' of Kim and Baldwin,[1] to those proposing that collapse proceeds folding, as in the three-stage random-search process suggested by Shakhnovich and co-workers,[2] with other models suggesting a concurrent collapse and formation of structure, as in the ''compaction-induced structure'' and the hydrophobic zipper models of Dill and co-workers.[3,4] (For reviews of these and other models, see Refs. 5–9.) Despite many experiments and computer simulations that provide anecdotal evidence supporting one model or another, the basic questions regarding the relationship between the processes that occur during folding remain unanswered, especially given the great uncertainties regarding the nature of the interactions that stabilize the native structure.

One way to approach these issues is from an evolutionary perspective. The need to fold puts strong constraints on the properties of the protein sequence, constraints that must be satisfied during the process of natural selection. We can consider how evolution might produce proteins that solve the protein-folding problem, and investigate the consequences of this ''solution.'' More directly, such an approach can help us design proteins with specific functions and properties that can correctly and consistently fold. For instance, Bryngelson and Wolynes approached the requirements for a protein to be able to fold rapidly using ideas adapted from spin-glass theory.[10–12] According to this model, the folding transition must compete with an alternative transition to a glassy state where movement on the free-energy landscape becomes slow and non-Arrhenius and nonergodic in the thermodynamic limit. As the glassy state represents the impossibility of finding the native state, the important parameter is the ratio of the folded transition temperature T_f to the glass transition temperature T_g. This ratio should presumably be large in order to ensure rapid folding. Wolynes and co-workers demonstrated that, in the context of a particularly simple model, this could be accomplished by having the free energy of the native state sufficiently low with respect to the distribution of the free energies of the non-native states.[13,14] Shakhnovich and co-workers demonstrated that such a condition was sufficient to provide fast-folding proteins during lattice simulations.[15–17]

In contrast, Thirumalai and co-workers argued for the importance of the temperature of compaction transition, T_θ, as an important characteristic of the protein thermodynamics.[18,19] According to their model, while proteins need a large value of T_f relative to T_g to fold quickly, T_f has to be lower than T_θ. Assuming that T_g and T_θ are both only sensitive to the generic features of amino acid sequences rather than the exact specifics of the individual sequence, one can maximize the ratio of T_f to T_g by minimizing the relative temperature gap between T_θ and T_f, eliminating the need to model the more difficult T_g.

Govindarajan and Goldstein used the spin-glass optimization method of Wolynes and co-workers to obtain the optimal set of interactions for every maximally compact conformation of a 27-residue protein confined to a $3\times3\times3$ cubic lattice.[20] In addition to ignoring the role of noncompact conformations and the interplay of compaction and folding, they were only able to derive relative energy parameters, as all maximally compact conformations have the same number of contacts. In this paper, we use a simple analytical model to extend the work of Govindarajan and Goldstein to include

[a)]Author to whom correspondence should be addressed. Phone: (313)763-8013; fax: (313)764-3323; electronic mail: richardg@umich.edu

 0021-9606/97/107(11)/4408/8/$10.00

noncompact states, allowing us to determine the average magnitude of the contact energies rather than just their relative values. We do this by taking advantage of recent work modeling folding as diffusion on a multidimensional free-energy landscape.[21,22] Because we are interested in understanding the relationship between compaction and formation of correct structure, our analytical theory is based on two different order parameters, one representing the degree of compactness, the other a measure of the progress towards the folded state. We first construct a one-dimensional reaction coordinate through the two-dimensional order parameter space, compute the free-energy and effective diffusion coefficient for motion along this reaction coordinate, and then model the folding as diffusion along this coordinate. This allows us to compute how the dynamics depends both on the temperature and the strength of the interactions. We find a situation somewhat more complicated than the analysis of Thirumalai indicates. In general, for marginally-stable proteins, the most rapid folding occurs with relatively strong average interactions. This is because such interactions reduce the entropy of the states competing with the folded state, increasing the value of T_f relative to T_g. Although T_f is increased, T_θ is more affected, resulting in a *large* difference between T_f and T_θ for the fastest-folding proteins.

II. THEORY

Wolynes and co-workers developed a description of protein folding as diffusion on the free-energy landscape.[10,11,21,22] This model is based on the parallel assumptions that folding occurs as the gradual buildup of structure, so that changes of the relevant order parameters are smooth and continuous, and that the time that the protein remains trapped in any individual state is short relative to the overall folding time. Under these conditions, the ensemble of states described by various values of the order parameters can be coarse-grained, resulting in a landscape that represents the free energy as a function of only these order parameters. Using this model, the evolution of a population of folding proteins can be described as a flux across the free-energy landscape, with changes in the order parameters characterized by a diffusion coefficient D which is a function of the roughness of the landscape. Folding times can then be easily computed using the diffusion equation. We apply this technique to measure how the folding rate depends upon the strength of nonspecific hydrophobic interactions relative to the sequence–dependent interactions that determine the final native state. By studying what conditions result in rapid folding, we can gain insight into how proteins that have evolved to fold quickly might actually fold, and how fast-folding artificial proteins can be designed.

We characterize a protein conformation by two order parameters: η, reflecting compactness, and ρ, a measure of the similarity to the native state. For both of these parameters, we focus on the number and types of contacts between residues, as motivated by earlier work showing that local propensities are probably not the major cause of stability.[23,24] η is defined as the total number of contacts divided by the number of contacts in the native state, which is assumed to be maximally compact. ρ is defined as the total number of native contacts divided by the number of contacts in the native state. ρ and η must satisfy the inequality $0 \leqslant \rho \leqslant \eta \leqslant 1$. We first derive expressions for the free-energy and diffusion coefficient as a function of the order parameters. We then construct an approximate reaction pathway through the conformational space, and solve the diffusion equation for a population of proteins folding from a random state to the folded state. The folding time is then computed for different values of the various interaction parameters.

The derivation of the free energy closely follows the procedure developed by Bryngelson and Wolynes,[12] based on Flory's theory of excluded volume.[25,26] The derivation begins by finding an expression for the entropy of an N-residue protein as a function of ρ, η, and energy E. We start with the expression for the number of states of an N-residue polymer with an end-to-end distance between R and $R+dR$, explicitly including the effect of excluded volume by multiplying a spherically weighted Gaussian distribution by the probability that any particular conformation with a given end-to-end distance would obey excluded volume,

$$\Omega_0(R)dR = C\nu^N \left(\frac{\sigma}{R}\right)^{3N} \frac{\Gamma[(R/\sigma)^3+1]}{\Gamma[(R/\sigma)^3-N+1]} \left(\frac{1}{R_0}\right) \times \left(\frac{R}{R_0}\right)^2 \exp\left[-\frac{3}{2}\left(\frac{R}{R_0}\right)^2\right] dR, \tag{1}$$

where C is a constant of order one, ν is the number of conformations for each residue, σ is the monomer radius and $R_0^2 = Nl^2$, where l is the average distance between monomers. We assume that $\sigma = l = 1$ in our model. Under these conditions the value of $\sqrt{\langle R^2 \rangle}$, the root-mean-squared value of the end-to-end distance, scales approximately as $N^{0.6}$, as would be expected for a polymer obeying excluded volume in a good solvent.[25]

We can put this expression in terms of η by assuming that the total number of contacts is approximately linear with the packing fraction, which varies with R to the negative third power

$$\eta = N\left(\frac{\sigma}{R}\right)^3. \tag{2}$$

Substituting this expression into Eq. (1) yields

$$\Omega_0(\eta)d\eta = \frac{1}{3}\,\nu^N \left(\frac{\eta}{N}\right)^N \frac{\Gamma\left(\frac{N}{\eta}+1\right)}{\Gamma\left(\frac{N}{\eta}-N+1\right)} N^{-1/2}\eta^{-2} \times \exp\left[-\frac{3}{2}\,(N^{-1/3}\eta^{-2/3})\right] d\eta. \tag{3}$$

We assume the number of native contacts in the folded state is equal to zN, where z is the average number of contacts per residue divided by two, and is roughly equal to unity. The

total number of contacts is then equal to $zN\eta$, while the number of native contacts is equal to $zN\rho$. The total number of states with $zN\rho$ native contacts and $zN(\eta-\rho)$ non-native contacts is simply equal to the total number of states with $zN\eta$ contacts times the probability that $zN\rho$ of these contacts are native, or

$$\Omega_0(\eta,\rho)=\Omega_0(\eta)\binom{z\eta N}{z\rho N}P_{\rm nat}^{z\rho N}P_{\rm non}^{z(\eta-\rho)N}. \tag{4}$$

$P_{\rm nat}$ is the probability of any formed contact being native and $P_{\rm non}=1-P_{\rm nat}$ is the corresponding probability of the contact being non-native. One advantage of using order parameters based on contacts is that we do not have to use any artificial constraints to prevent unphysical folded yet extended states.[12] Instead, an additional complication is caused by the fact that the probability of any contact being native will depend upon how many other native contacts there are, in that native contacts constrain the protein from making other non-native contacts, a cause of much of the cooperativity in protein folding. We assume a simple quadratic form for this dependence, and write

$$P_{\rm nat}=P_{\rm nat}^0(1+\alpha\rho^2), \tag{5}$$

where $P_{\rm nat}^0$ is the probability of a contact being native in the limit of $\rho\to 0$, assumed to be equal to the number of possible native contacts divided by the total number of possible contacts, and α is a constant that will be determined by the need for the conformational entropy of the completely folded state ($\eta=1,\rho=1$) to be zero. Using this approach, we are able to model the cooperativity through the calculation of the entropy, rather than having to postulate cooperative energetic interactions.

Combining Eqs. (4) and (5) yields

$$\Omega_0(\eta,\rho)=\Omega_0(\eta)\,\frac{(zN\eta)![P_{\rm nat}^0(1+\alpha\rho^2)]^{zN\rho}[1-P_{\rm nat}^0(1+\alpha\rho^2)]^{zN(\eta-\rho)}}{[zN(\eta-\rho)]!(zN\rho)!}. \tag{6}$$

The next step is to include the effect of intramolecular interactions by considering how many of these states have a given energy E. Again following Bryngelson and Wolynes, we start by using the central limit theorem to represent the probability of a polymer chain having total attractive energy E as a Gaussian with an average $\overline{E}$ and standard deviation ΔE that are functions of η and ρ.[12]

$$P(E|\eta,\rho)=\frac{1}{[2\pi\Delta E(\eta,\rho)^2]^{1/2}}\exp\left\{-\frac{[E-\overline{E}(\eta,\rho)]^2}{2\Delta E(\eta,\rho)^2}\right\}, \tag{7}$$

where $P(E|\eta,\rho)$ is the conditional probability that a protein with given values of the order parameters η and ρ would have energy E. This results in the final expression

$$\Omega(\eta,\rho,E)=\Omega_0(\eta,\rho)P(E|\eta,\rho)=\frac{1}{3}\,\nu^N\left(\frac{\eta}{N}\right)^N\frac{\Gamma\left(\frac{N}{\eta}+1\right)}{\Gamma\left(\frac{N}{\eta}-N+1\right)}N^{-1/2}\eta^{-2}\exp\left[-\frac{3}{2}\,(N^{-1/3}\eta^{-2/3})\right]$$
$$\times\frac{(zN\eta)![P_{\rm nat}^0(1+\alpha\rho^2)]^{zN\rho}[1-P_{\rm nat}^0(1+\alpha\rho^2)]^{zN(\eta-\rho)}}{[zN(\eta-\rho)]!(zN\rho)!}\,\frac{1}{[2\pi\Delta E(\eta,\rho)^2]^{1/2}}\exp\left\{-\frac{[E-\overline{E}(\eta,\rho)]^2}{2\Delta E(\eta,\rho)^2}\right\}. \tag{8}$$

In order to convert from a microcanonical ensemble with a specified energy E to a canonical ensemble with a temperature T, we use the Legendre transformation

$$\frac{1}{T}=\left(\frac{\partial S}{\partial E}\right), \tag{9}$$

where $S(\eta,\rho,E)=\ln\Omega(\eta,\rho,E)$ to yield[12]

$$E=\overline{E}(\eta,\rho)-\frac{\Delta E(\eta,\rho)^2}{T}. \tag{10}$$

Substituting Eq. (10) into Eq. (8) yields the temperature dependence of the entropy

$$S(T,\eta,\rho)=S_0(\eta,\rho)-\frac{\Delta E(\eta,\rho)^2}{2T^2}, \tag{11}$$

where $S_0(\eta,\rho)=\log(\Omega_0(\eta,\rho))$.

Combining Eqs. (10) and (11), we finally arrive at the expression for free energy as a function of temperature, η, and ρ

$$F(T,\eta,\rho)=E-TS=\overline{E}(\eta,\rho)-\frac{\Delta E(\eta,\rho)^2}{2T}-TS_0(\eta,\rho). \tag{12}$$

We are interested in understanding the relative importance of nonspecific hydrophobic forces of compaction and side-chain specific interactions in directing the folding process. For this reason, we model the energies of random interactions as forming a distribution with average interaction energy γ_0 and standard deviation $\Delta\gamma$. γ_0 then represents these general forces of compaction, while the magnitude of

$\Delta\gamma$ characterizes the range of interaction strengths based on the properties of the individual residues. Following the ''principle of minimal frustration,''[10] we would expect the variation among the contact energies of native contacts to be significantly smaller than the variation in the energies of random contacts. In fact, lattice simulations have indicated that homogeneity of native interactions increases a protein's ability to fold rapidly.[27] We neglect the variation in strengths of contacts formed in the native state, and assume that these contacts have the same interaction strength, given by $\gamma_N + \gamma_0$, where γ_N represents the stabilizing strength of a native contact relative to the average of the non-native contacts. With $zN\rho$ native contacts and $zN(\eta-\rho)$ non-native contacts, the values of $E(\eta,\rho)$ and $\Delta E(\eta,\rho)$ are given by

$$\overline{E(\eta,\rho)} = zN(\eta\gamma_0 + \rho\gamma_N) \tag{13}$$

and

$$\Delta E(\eta,\rho)^2 = zN(\eta-\rho)\Delta\gamma^2. \tag{14}$$

We can characterize the free-energy landscape by looking at the types of minima that exist. At large temperatures, the entropic terms dominate, and the protein will exist predominantly in an unfolded (small η and ρ) state. At low temperatures, the dominant thermodynamic state will be the folded (large η and ρ) state. For large negative values of γ_0 there is a range of temperature where a compact but unfolded state (large η, small ρ) is thermodynamically dominant. Such a state might correspond to the molten-globule state observed either transiently or at equilibrium in acidic conditions in the presence of denaturants. We can also determine which values of the parameters result in a free-energy barrier between the unfolded and folded states. If such a barrier does not exist, folding would be diffusion limited, corresponding to the ''type 0'' folding described by Wolynes and co-workers.[28] Folding under conditions where a barrier *does* exist would correspond to ''type I'' or ''type II'' folding. There is some degree of disorder even in folded proteins, and the minimum in the free energy corresponding to the folded state is generally not at $\eta=\rho=1$. We define all states with $\rho\geqslant 0.8$ as ''folded,'' while all states with $\eta\geqslant 0.5$ are defined as ''compact.''

In order to connect more directly with the theory developed by Govindarajan and Goldstein, we use parameter values representing those appropriate for a 27-residue protein confined to a cubic lattice. For a cubic lattice, $\nu=5$. The maximally compact conformations make 28 contacts, so $z=28/27$. With a total of 156 possible contacts between nonadjacent residues, $P^0_{\rm nat}=28/156$. Setting the value of the conformational entropy at $\rho=\eta=1$ to zero sets α in Eq. (5) to 2.703. As all energies are relative, we set the width of the distribution of random compact states ($\Delta E(\eta=1,\ \rho=0) = \sqrt{zN}\Delta\gamma$) equal to one, or $\Delta\gamma=0.19$. We set the value of $\gamma_N=-0.45$ based on the work of Govindarajan and Goldstein, which suggested that the largest possible value of T_f/T_g would occur when the the stability of the folded state relative to compact unfolded states ($zN\gamma_N/\sqrt{zN}\Delta\gamma$) is equal to the square root of the number of possible contacts.[20,23]

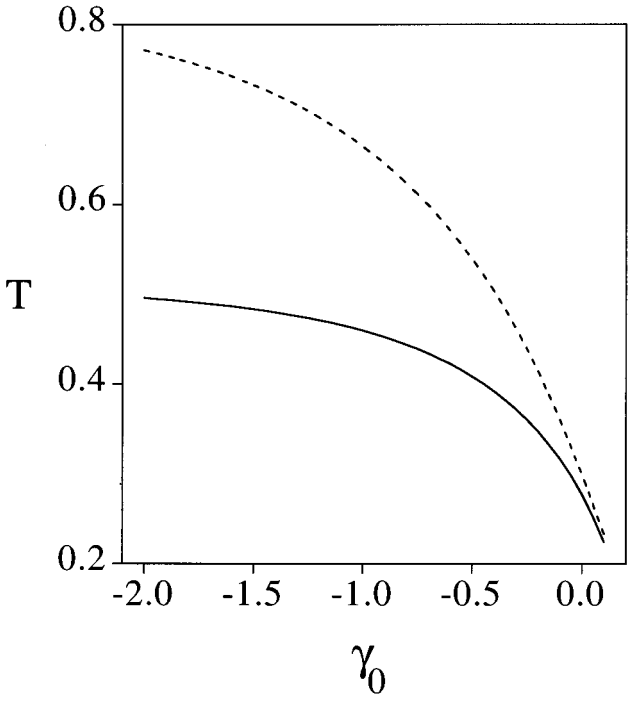

FIG. 1. Folding temperature T_f (—) and compaction transition temperature T_θ (---) as a function of the average interaction γ_0.

The folding temperature T_f is defined as the temperature where the Boltzmann-weighted sum of folded states is equal to half of the total Boltzmann-weighted sum. Similarly the compaction transition temperature T_θ represented the temperature where the Boltzmann-weighted sum of compact states represents half of the total population. The dependence of T_f and T_θ as a function of γ_0 for the parameter values described above are shown in Fig. 1.

We are interested in studying the transition from the unfolded state to folded state, with the order parameters η and ρ changing as the protein compacts and folds. In order to simplify the computation, we first derive a reaction coordinate by computing an approximate reaction pathway. The totally unfolded initial state is modeled as the state of maximum entropy. We then find the saddle points in the free-energy landscape, and calculate an approximate reaction pathway by calculating the trajectories of steepest slope from the unfolded state and all of the saddle points to the closest minima. The various piece-wise continuous segments are then assembled into a continuous reaction pathway. Movement along this path is associated with the reaction coordinate ξ, with the value of ξ corresponding to the folded state ($\rho=0.8$) defined as ξ_F. This approach is in the same spirit of the concept of minimum-energy path in chemical kinetics.[29] Typical reaction pathways for $T/T_f=0.8$ and various values of γ_0 are shown in Fig. 2.

Bryngelson and Wolynes demonstrated that under certain assumptions the diffusion coefficient for changes of the reaction coordinates in a model such as ours can be represented with the Ferry-law expression

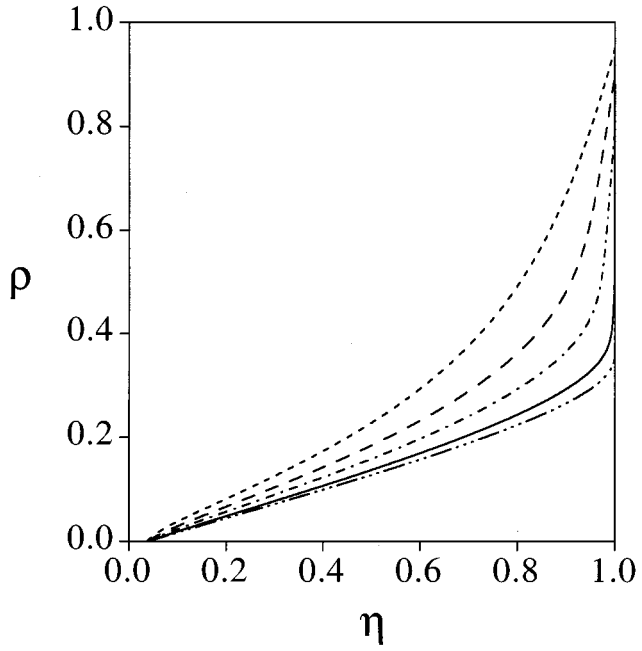

FIG. 2. Folding pathway showing changes in the order parameters η and ρ corresponding to movement along the reaction coordinate ξ, for $T/T_f=0.8$ and various values of γ_0: $\gamma_0=-0.2$ (---), $\gamma_0=-0.4$ (– – –), $\gamma_0=-0.6$ (·-·-·-), $\gamma_0=-0.85$ (optimal condition) (—), and $\gamma_0=-1.0$ (—···—···).

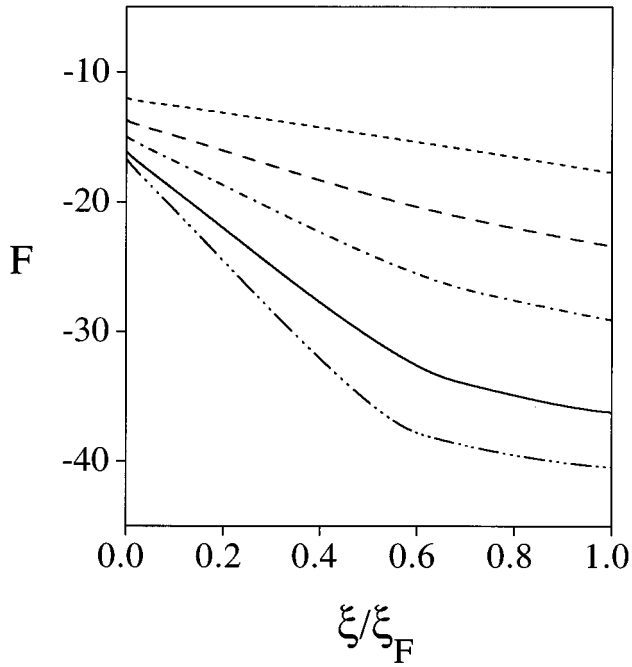

FIG. 3. Free-energy profile along the folding pathway for $T/T_f=0.8$ and various values of γ_0: $\gamma_0=-0.2$ (---), $\gamma_0=-0.4$ (– – –), $\gamma_0=-0.6$ (·-·-·-), $\gamma_0=-0.85$ (optimal condition) (—), and $\gamma_0=-1.0$ (—···—···). ξ_F represents the value of the reaction coordinate for the folded state ($\rho=0.8$).

$$D(\eta,\rho,T)=D_0\ \exp\left[-\frac{\Delta E^2(\eta,\rho)}{(kT)^2}\right], \tag{15}$$

where D_0 is related to the time-scale for conformational changes of the protein.[11] The conceptual basis of the super-Arrhenius behavior is that as the temperature decreases, the system is more likely to populate lower energy and thus deeper local minima, reducing the ability of the system to escape these minima more than would be predicted with a simple Arrhenius law. This also assumes that the protein is sufficiently optimized for folding to avoid glassy behavior. We use this expression in our model, although the assumption made by Bryngelson and Wolynes, that the transition states representing escape from local traps have an energy distribution equivalent to the ensemble of states for the protein with that value of ρ and η becomes questionable as the strength of the contacts increases. We assume that this diffusion coefficient is appropriate for motion along the reaction coordinate ξ, whether this involves portions of the reaction pathway involving predominantly compaction (increasing η) or folding (increasing ρ). The free-energy profile and diffusion coefficient along the folding pathway for $T/T_f=0.8$ and various values of γ_0 are shown in Figs. 3 and 4. As we are only interested in relative rates, we do not attempt to estimate D_0.

Once the free-energy landscape has been constructed, the reaction pathway defined, and the diffusion coefficient calculated, we can use a simple diffusion equation to model the population of the folding proteins $P(\xi,t)$[22]

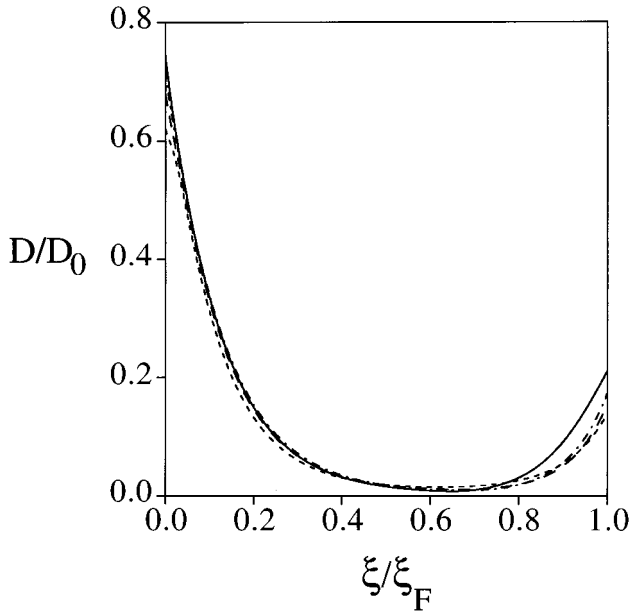

FIG. 4. Relative diffusion coefficient D/D_0 along the folding pathway for $T/T_f=0.8$ and various values of γ_0: $\gamma_0=-0.2$ (---), $\gamma_0=-0.4$ (– – –), $\gamma_0=-0.6$ (·-·-·-), $\gamma_0=-0.85$ (optimal condition) (—), and $\gamma_0=-1.0$ (—···—···).

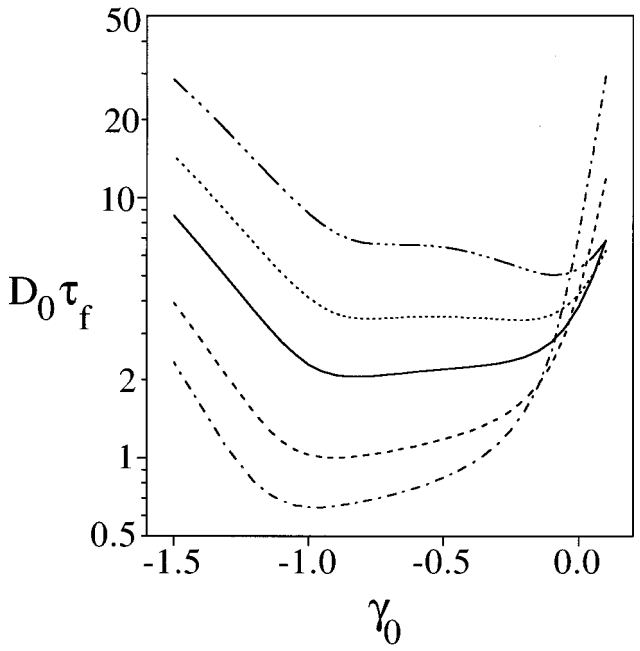

FIG. 5. Folding time τ_f as a function of γ_0 for various values of T/T_f: $T/T_f=0.7$ (—···—···), $T/T_f=0.75$ (···), $T/T_f=0.8$ (—), $T/T_f=0.9$ (---), and $T/T_f=1.0$ (·-·-·). Folding times are scaled by the unknown diffusion coefficient D_0.

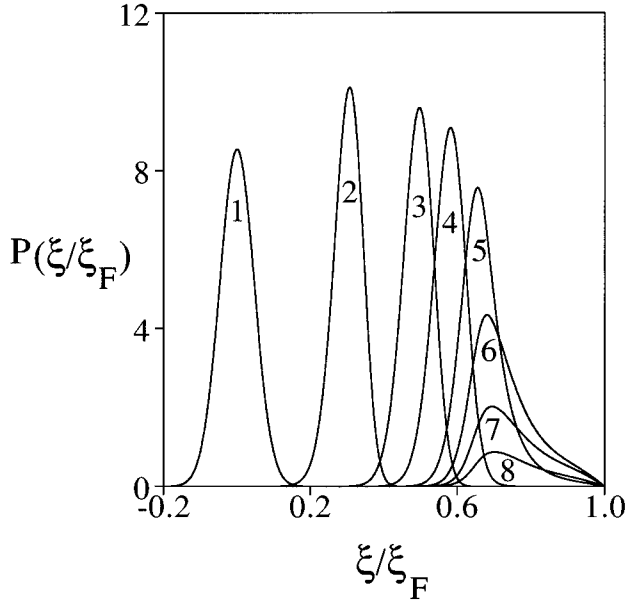

FIG. 6. Time evolution of the population of proteins along the reaction pathway $P(\xi/\xi_F)$ for $T/T_f=0.8$ and $\gamma_0=-0.85$. (1) $t=0$; (2) $t=0.05/D_0$; (3) $t=0.25/D_0$; (4) $t=0.50/D_0$; (5) $t=1.00/D_0$; (6) $t=1.50/D_0$; (7) $t=2.00/D_0$; (8) $t=2.50/D_0$.

$$\frac{\partial P(\xi,t)}{\partial t}=\frac{\partial}{\partial \xi}\left\{D(\xi)\left[\frac{\partial P(\xi,t)}{\partial \xi}+\frac{P(\xi,t)}{kT}\frac{\partial F(\xi)}{\partial \xi}\right]\right\}. \quad (16)$$

The initial population is set equal to a narrow Gaussian distribution of ξ values centered around the point of maximum entropy. $P(\xi)$ is set equal to zero at the value of ξ where $\rho=0.8$, as folded proteins are removed from the population. The length of time necessary for the remaining population of nonfolded proteins to decrease to $1/e$ of their original population is defined as the folding time τ_f. Because of the boundary conditions at $\rho=0.8$, this time reflects the mean first passage time. Similarly, the time required for the population of noncompact proteins to decrease to $1/e$ of their original value is defined as the compaction time τ_θ.

III. RESULTS

In general, for temperatures around the temperature of the folding transition, increasing the temperature generally decreases the folding time, providing one more possible explanation why proteins are only marginally stable. The value of T/T_f is then determined by requirements other than the need to fold quickly, such as the necessity for the unfolded state to be sufficiently thermodynamically inaccessible so that the protein is not overly susceptible to proteolysis. We are then interested in trying to find the optimal value of γ_0 when T/T_f is determined by such thermodynamic criteria. Figure 5 displays the main results of this paper, the folding time τ_f as a function of γ_0 for various values of T/T_f.

For marginally stable proteins (with T/T_f between 0.8 and 1.0), the optimal values of γ_0 are in the range of -0.85 to -0.95, indicating that the magnitude of the optimal average interaction between two residues is quite large relative to the variation in random non-native interactions ($\Delta\gamma=0.19$) and the strength of the native interactions relative to the non-native interactions ($\gamma_N=-0.45$). A typical time evolution of $P(\xi)$ for $T/T_f=0.8$ and γ_0 equal to its optimal value of -0.85 is shown in Fig. 6. We can similarly plot the relative population of unfolded ($\eta<0.5$), compact but unfolded ($\eta\geqslant 0.5,\rho<0.8$) and folded population ($\rho\geqslant 0.8$) as a function of time, as shown in Fig. 7. Under optimal conditions, the protein quickly contracts, forming a relatively compact state with a compaction time $\tau_\theta=0.06/D_0$, followed by a slower folding with $\tau_f=2.05/D_0$. The fastest folding corresponds, unsurprisingly, to type 0 folding, where there is no free-energy barrier and the rate of folding is diffusion limited.[28] The compact state with a low amount of correct structure corresponds to the smallest value of the diffusion coefficient, as ΔE^2 is proportional to $\eta-\rho$ in this simple model. It is this compact unfolded state that corresponds to the ''kinetic bottleneck.'' Our results suggest that under optimal conditions, proteins would collapse quickly to a relatively compact state with limited correct tertiary structure, similar to what has been postulated to exist in the so-called ''molten-globule state.'' Such results have been observed both experimentally[30–34] and in computer simulations.[2,35–40]

As the T/T_f is reduced below 0.8, the dependence of the folding time on γ_0 becomes more complicated. In fact, at $T/T_f=0.75$ there are two different optimal values of γ_0 with almost identical values of τ_f; $\gamma_0=-0.75$, where the dynamics resemble that discussed above, with fast compaction followed by slower folding, and $\gamma_0=-0.20$, where compaction and folding proceed more simultaneously. At even lower val-

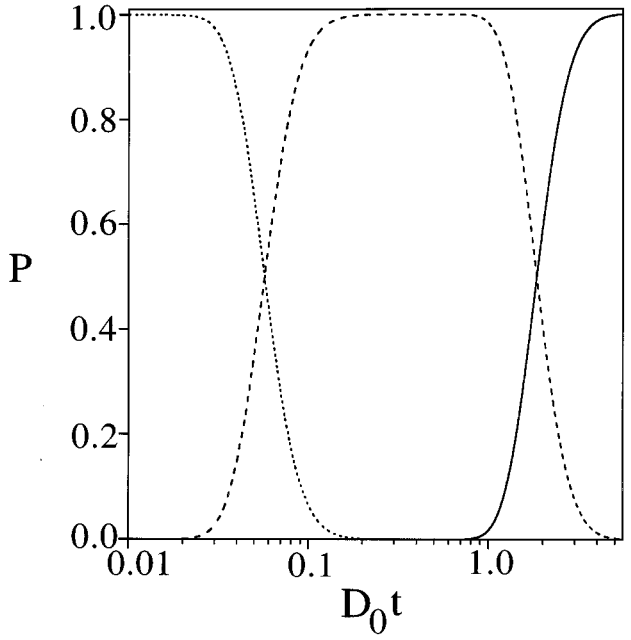

FIG. 7. Population of unfolded state (···), compact unfolded state (---) and folded state(—) as a function of time for $T/T_f=0.8$ and $\gamma_0=-0.85$. Times, shown on a log scale, are multiplied by the unknown diffusion coefficient D_0.

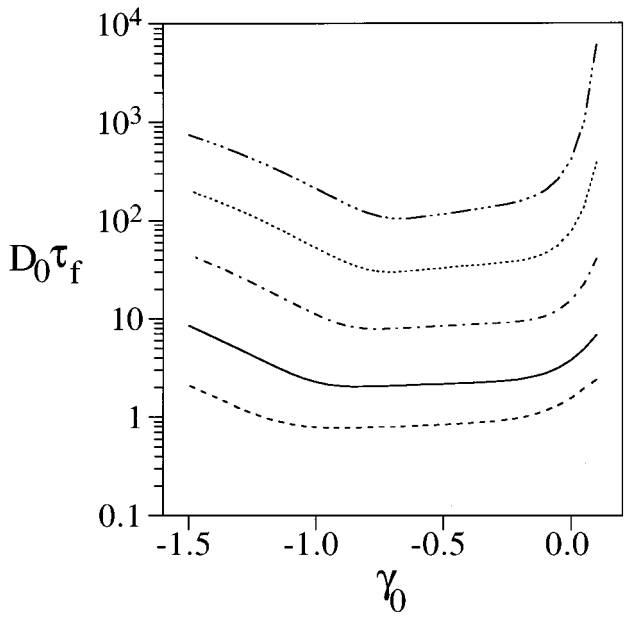

FIG. 8. Folding time τ_f as a function of γ_0 for various values of γ_N, with T/T_f set equal to 0.8: $\gamma_N=-0.35$ (—···—···), $\gamma_N=-0.37$ (···), $\gamma_N=-0.40$ (·-·-·), $\gamma_N=-0.45$ (—), and $\gamma_N=-0.50$ (---). Folding times are scaled by the unknown diffusion coefficient D_0.

ues of T/T_f, faster folding is observed with moderately negative values of γ_0.

In the above plots, we fixed γ_N at the approximate maximum value consistent with the lattice model. We can also explore how the dependence of the folding time on γ_0 is affected by the value of γ_N. As shown in Fig. 8, the general results of the theory are not overly affected by changes in this parameter.

IV. DISCUSSION

We can understand the existence of an optimal value of γ_0 by considering all of the various factors that change as γ_0 is modified. There are three major factors. As the compact folded state has more contacts than the unfolded state, making γ_0 increasingly negative increases the overall gradient of the free energy, as can be seen in Fig. 3. This effect of this change is modulated by the fact that for larger negative values of γ_0, the folding pathway proceeds with an initial rapid compaction, and then through the generation of correct contacts maintaining a roughly constant degree of compaction. As a result, the free-energy gradient for the latter, slower parts of protein folding is actually reduced. A second effect is due to the fact that as the protein compacts more quickly, the protein folding pathway passes through regions of increased $\eta-\rho$, where ΔE is a maximum and the diffusion coefficient is a minimum. The major effect, however, is due to the change in T_f with γ_0, as shown in Fig. 1. As γ_0 is made increasingly negative, the ensemble of unfolded states becomes more and more dominated by compact states, and the entropy of the unfolded state decreases. Folding can then occur at a higher temperature. This means that the diffusion coefficient can be larger, even with a reaction pathway that proceeds through increasingly rougher areas of the free-energy landscape. As a result, for marginally stable proteins, the folding time decreases with increasingly negative γ_0, until γ_0 is large enough so that there is little further change in T_f, as shown in Fig. 1. It is only under these circumstances that the other effects start to become more dominant, and the folding time increases with further changes in γ_0.

Thirumalai argued that fast folding sequences have small values of $\sigma=(T_\theta-T_f)/T_\theta$.[18,19] We find that the situation is a bit more complicated. While making γ_0 increasingly negative increases T_f, it has a much larger effect on T_θ. The result is that the large negative values of γ_0 under optimal folding conditions correspond to quite large values of σ. That is not to say that our results directly contradict the results of Thirumalai. If γ_0 is mostly determined by the generic properties of amino acids, it still may be true that γ_0 is roughly a constant for all biologically relevant sequences. Under these more specific conditions, there may be more of a connection between σ and the folding time.

It is interesting to note that the results of this simple model so closely resemble experimental observations of protein folding. We find that under optimal folding conditions, a state similar to a molten-globule state results, where there is compaction and a relatively small amount of tertiary structure. Folding then involves the search for correct contacts within this ensemble of compact structures. These results are consistent with models such as the three-stage random-search model proposed by Shakhnovich and co-workers for small proteins.[2] (The effect of nucleation, possibly important in the folding of larger proteins,[17] cannot be incorporated into the current model.) These results also suggest that the

structure is formed by interactions that would be particularly relevant in a compact state, such as patterns of hydrophobicity.[7] This result is consistent with observations that binary encodings of hydrophobicity are often sufficient to stabilize protein structures, and are frequently observed in sequences of biological proteins,[41,42] and that hydrophobicity is conserved during evolution much more than local propensities.[24] This also may indicate the importance of hydrogen bond formation in determining the folded structure, as once the protein collapsed and much of the protein was inaccessible to solvent, the stabilizing effect of intraprotein hydrogen bonds would not be cancelled by the loss of protein-solvent hydrogen bonds.

For proteins with high stability (T/T_f less than 0.75), optimal values of γ_0 are much closer to zero, and the optimal folding involves simultaneous compaction and formation of native contacts. This suggests that the optimal folding situation may be dependent on the thermodynamic properties of the protein, and that different proteins may have evolved following different folding strategies.

Finally, these results suggest that simulations which neglect noncompact states might yield reasonable results. Conversely, the complicated dependence of the folding rate on γ_0 and the dependence of this effect on T demonstrate the danger of making general conclusions based on simulations performed under only a limited number of values of this parameter.

ACKNOWLEDGMENTS

We would like to thank Kurt Hillig and James Raines for computational assistance, and Sridhar Govindarajan for helpful discussions. Financial support was provided by the College of Literature, Science, and the Arts, the Program in Protein Structure and Design, the Horace H. Rackham School of Graduate Studies, NIH Grant LM0577, and NSF equipment Grant BIR9512955.

[1] K. Kim and R. Baldwin, Annu. Rev. Biochem. **51**, 459 (1982).
[2] A. Šali, E. I. Shakhnovich, and M. J. Karplus, Nature (London) **369**, 248 (1994).
[3] K. A. Dill, Biochem. **24**, 1501 (1985).
[4] K. A. Dill, K. M. Fiebig, and H. S. Chan, Proc. Nat. Acad. Sci., U.S.A. **90**, 1942 (1993).
[5] N. Go, Annu. Rev. Biophys. Bioeng. **12**, 183 (1983).
[6] M. Karplus and E. Shakhnovich, Protein folding: Theoretical studies, in *Protein Folding*, edited by T. Creighton, (Freeman, New York, 1992), pp. 127–195.
[7] K. A. Dill *et al.* Protein Sci. **4**, 561 (1995).
[8] M. Karplus and A. Šali, Current Biology **5**, 58 (1995).
[9] P. G. Wolynes, Z. Luthey-Schulten, and J. N. Onuchic, Current Biology **3**, 425 (1996).
[10] J. D. Bryngelson and P. G. Wolynes, Proc. Nat. Acad. Sci. U.S.A. **84**, 7524 (1987).
[11] J. D. Bryngelson and P. G. Wolynes, J. Phys. Chem. **93**, 6902 (1989).
[12] J. D. Bryngelson and P. G. Wolynes, Biopolymers **30**, 171 (1990).
[13] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, Proc. Nat. Acad. Sci., U.S.A. **89**, 4918 (1992).
[14] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, Proc. Nat. Acad. Sci., U.S.A. **89**, 9029 (1992).
[15] A. Šali, E. I. Shakhnovich, and M. J. Karplus, J. Mol. Biol. **235**, 1614 (1994).
[16] E. I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).
[17] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, Biochem. **33**, 10026 (1994).
[18] D. Thirumalai, J. Phys. I France. **5**, 1457 (1995).
[19] D. K. Klimov and D. Thirumalai, Phys. Rev. Lett. **76**, 4070 (1996).
[20] S. Govindarajan and R. A. Goldstein, Biopolymers **36**, 43 (1995).
[21] J. N. Onuchic, P. G. Wolynes, Z. Luthey-Schulten, and N. D. Socci, Proc. Nat. Acad. Sci., U.S.A. **92**, 3626 (1995).
[22] N. D. Socci, J. N. Onuchic, and P. G. Wolynes, J. Chem. Phys. **104**, 5860 (1996).
[23] S. Govindarajan and R. A. Goldstein, Proteins **22**, 413 (1995).
[24] J. M. Koshi and R. A. Goldstein, Proteins **27**, 336 (1997).
[25] P. J. Flory, J. Chem. Phys. **17**, 303 (1949).
[26] I. C. Sanchez, Macromolecules **12**, 980 (1979).
[27] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, Folding and Design **1**, 221 (1996).
[28] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins **21**, 167 (1995).
[29] K. J. Laidler, *Chemical Kinetics*, 3rd ed. (Harper & Row, New York, 1987).
[30] A. F. Chaffotte, Y. Guillou, M. Delepierre, H.-J. Hinz, and M. E. Goldberg, Biochem. **30**, 8067 (1991).
[31] A. F. Chaffotte, C. Cadieux, Y. Guillou, and M. E. Goldberg, Biochem. **31**, 4303 (1992).
[32] L. S. Itzhaki, P. A. Evans, C. M. Dobson, and S. E. Radford, Biochem. **33**, 5212 (1994).
[33] V. R. Agashe, M. C. R. Shastry, and J. B. Udgaonkar, Nature **377**, 754 (1995).
[34] W. A. Houry, D. M. Rothwarf, and H. A. Scheraga, Biochem. **35**, 10125 (1996).
[35] E. I. Shakhnovich, G. Farztdinov, A. M. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).
[36] C. J. Camacho and D. Thirumalai, Phys. Rev. Lett. **71**, 2505 (1993).
[37] C. J. Camacho and D. Thirumalai, Proc. Nat. Acad. Sci., U.S.A. **90**, 6369 (1993).
[38] M. Fukugita, D. Lancaster, and M. G. Mitchard, Proc. Nat. Acad. Sci., U.S.A. **90**, 6365 (1993).
[39] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **101**, 1519 (1994).
[40] A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich, Biochem. **34**, 3066 (1995).
[41] S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, and M. H. Hecht, Science **262**, 1680 (1993).
[42] M. W. West and M. H. Hecht, Protein Sci. **4**, 2032 (1995).