Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties

Thomas Veitshans[1], Dmitri Klimov[2] and Devarajan Thirumalai[2]

Background: Recent experimental and theoretical studies have revealed that protein folding kinetics can be quite complex and diverse depending on various factors such as size of the protein sequence and external conditions. For example, some proteins fold apparently in a kinetically two-state manner, whereas others follow complex routes to the native state. We have set out to provide a theoretical basis for understanding the diverse behavior seen in the refolding kinetics of proteins in terms of properties that are intrinsic to the sequence.

Results: The folding kinetics of a number of sequences for off-lattice continuum models of proteins is studied using Langevin simulations at two different values of the friction coefficient. We show for these models that there is a remarkable correlation between folding time, τ_F, and $\sigma = (T_\theta - T_F)/T_\theta$, where T_θ and T_F are the equilibrium collapse and folding transition temperatures, respectively. The microscopic dynamics reveals that several scenarios for the kinetics of refolding arise depending on the range of values of σ. For relatively small σ, the chain reaches the native conformation by a direct native conformation nucleation collapse (NCNC) mechanism without being trapped in any detectable intermediates. For moderate and large values of σ, the kinetics is described by the kinetic partitioning mechanism, according to which a fraction of molecules Φ (kinetic partition factor) reach the native conformation via the NCNC mechanism. The remaining fraction attains the native state by off-pathway processes that involve trapping in several misfolded structures. The rate-determining step in the off-pathway processes is the transition from the misfolded structures to the native state. The partition factor Φ is also determined by σ: the smaller the value of σ, the larger is Φ. The qualitative aspects of our results are found to be independent of the friction coefficient. The simulation results and theoretical arguments are used to obtain estimates for timescales for folding via the NCNC mechanism in small proteins, those with less than about 70 amino acid residues.

Conclusions: We have shown that the various scenarios for folding of proteins, and possibly other biomolecules, can be classified solely in terms of σ. Proteins with small values of σ reach the native conformation via a nucleation collapse mechanism and their energy landscape is characterized by having one dominant native basin of attraction (NBA). On the other hand, proteins with large σ get trapped in competing basins of attraction (CBAs) in which they adopt misfolded structures. Only a small fraction of molecules access the native state rapidly when σ is large. For these sequences, the majority of the molecules approach the native state by a three-stage multipathway mechanism in which the rate-determining step involves a transition from one of the CBAs to the NBA.

Addresses: [1]Laboratoire de Spectrométrie Physique, associé au CNRS, Université J Fourier, Grenoble I, BP 87, 38402 Saint-Martin d'Hères Cedex, France. [2]Institute for Physical Science and Technology and Department of Chemistry and Biochemistry, University of Maryland, College Park, Maryland 20742, USA.

Correspondence: Devarajan Thirumalai
e-mail: thirum@ipst.umd.edu

Key words: collapse and folding transition temperature, kinetic partitioning mechanism, native conformation nucleation collapse, protein folding, three-stage multipathway mechanism

Received: 16 Sep 1996
Revisions requested: 14 Oct 1996
Revisions received: 01 Nov 1996
Accepted: 04 Nov 1996

Published: 16 Dec 1996
Electronic identifier: 1359-0278-002-00001

Folding & Design 16 Dec 1996, 2:1–22

Introduction

It has become clear over the past few years that the study of minimal models has given rise to a novel theoretical understanding of the kinetics of protein folding [1–8]. The general scenarios that have emerged from these studies are starting to be confirmed experimentally [9–15]. In particular, there is now some experimental support [11,14] for the kinetic partitioning mechanism (KPM) first described using minimal off-lattice models [5,16,17]. The principles emerging from these studies have also been used to predict the folding pathways and the nature of kinetic intermediates in specific proteins. For example, it was shown that the single disulfide intermediate 14–38 in bovine pancreatic trypsin inhibitor, which denotes that the structure of this intermediate contains a covalent disulfide bond between cysteines at location 14 and 38, forms early and decays before other more stable single intermediates start to form [18]. This theoretical prediction has subsequently been

verified experimentally [19]. Theoretical studies [1–8,20] have in fact provided, perhaps for the first time, a firm basis for understanding and predicting the overall scenarios that can arise in *in vitro* refolding kinetics of proteins. Since the general scenarios for refolding kinetics have been understood from a qualitative viewpoint, it is of interest to correlate in a quantitative manner the dependence of folding times for a number of sequences in terms of parameters that can be measured experimentally. In a recent paper [21], theoretical arguments were used to provide quantitative estimates of some of the important timescales that arise naturally according to the KPM. In this paper, we use computational studies to complement the previous work. We should note that Onuchic *et al.* [22] have also initiated complementary approaches to understand in a quantitative manner the folding kinetics of small α-helical proteins.

In our earlier work [23,24], we showed using lattice models that the foldability of proteins (i.e. the ability of a sequence with a unique native state to access it in finite timescales under folding conditions) can be understood in terms of two characteristic thermodynamic temperatures which are intrinsic to the sequence: T_θ, the collapse transition temperature, at which there is a transition from a random coil to an almost compact state, and the folding transition temperature, T_F, below which the polypeptide chain is predominantly in the native conformation. In the biochemical literature, T_F is roughly the melting temperature, T_m. It was established for a variety of sequences (in both two and three dimensions) that the folding time correlates extremely well with σ [23,24], where:

$$\sigma = \frac{T_\theta - T_F}{T_\theta} \qquad (1)$$

Based on very general arguments [21], it can be shown that $T_F \leq T_\theta$, hence $0 \leq \sigma \leq 1$. Both T_F and T_θ are sensitive functions of not only the sequence but also the external conditions. This can be verified experimentally and data in the literature in fact support this obvious result. For example, Alexander *et al.* [25] have shown for the IgG-binding protein that T_F (in their notation T_m) varies linearly with pH. These authors have also determined T_θ for two forms of IgG-binding protein. Thus, foldability of sequences and the associated kinetics for a given sequence can be altered by changing the external conditions. The major purpose of this article is to explore the folding kinetics as a function of σ using off-lattice simplified models of polypeptide chains. In addition, we provide detailed analysis of folding kinetics for a number of sequences at two values of the friction coefficient to assess the role of viscosity on the qualitative aspects of the folding scenarios.

The physical reason for expecting that σ would control the folding rates in proteins is the following. If σ is small, then $T_\theta \approx T_F$ and hence all the conformations that are sampled at $T \lesssim T_F$ have relatively high free energy. Any barrier that may exist between these high free energy mobile conformations can be overcome easily provided the temperature at which folding occurs is not too low. Thus, for small σ one can in principle fold a polypeptide chain at a relatively high temperature (in the range where collapse and the acquisition of the native conformation are almost synchronous) and access the native conformation rapidly. For these cases, the folding process would appear to be kinetically two-state-like [26]. Furthermore, studies based on lattice models suggest that for sequences with small and moderate values of σ, the kinetic accessibility of the native conformation together with its thermodynamic stability can be achieved over a relatively broad temperature range [23]. On the other hand, when $\sigma \approx 1$, then $T_F << T_\theta$, and in this case the folding process would inevitably be affected by kinetic traps and misfolded structures. Since some of these misfolded structures can have many elements in common with the native structure, they can be fairly stable [16,17]. Since T_F is low for sequences with large σ, these stable structures could have long lifetimes even if the free energy barriers separating them and the native state are only moderate. Thus, it is likely that sequences with $\sigma \approx 1$ are in general not foldable on biologically relevant timescales. These expectations are borne out in this study and a quantitative relationship between folding rates and σ is given. The energy landscape perspective can be used to argue that small values of σ correspond to the native state having a large native basin of attraction (NBA) [27] or funnel [2,3,28].

Before we close this introduction, a brief comment on the use of minimal models to understand folding kinetics is pertinent. This is especially important because their utility in getting insights into protein folding kinetics has been questioned [29]. The minimal models do not explicitly contain all the features that are known to be important in imparting stability to proteins, but many aspects of them do mimic the dominant interactions in proteins [1]. These involve chain connectivity, hydrophobicity as the driving force, and sequence heterogeneity. In addition, off-lattice models studied in this paper and elsewhere [30,31] which use a realistic representation of the potentials for α-carbons of a polypeptide chain yield (ϕ,ψ) values consistent with the Ramachandran plot [32]. The aspects of real proteins that are not faithfully represented here are sidechains and hydrogen bonds. Straub and Thirumalai (unpublished data) have argued that lower-order effects such as stability arising from hydrogen bonds are included in the simplified off-lattice models of the sort considered here in a coarse-grained manner. This is achieved by suitably renormalizing the dihedral angle potentials. Despite these important limitations, the studies based on minimal models of proteins have been the only source of concrete testable theoretical predictions in the field of protein folding kinetics [1–8]. Insights based on the energy landscape picture of folding have led, for example, to the microscopic picture of

native conformation nucleation collapse (NCNC) mechanism in refolding of proteins [16,17,20]. Recently, experimentalists have begun interpreting their data on certain proteins [15,33] using the concept of NCNC. Thus, despite certain limitations, these studies have already offered considerable insights into the folding kinetics of biomolecules [1–3,8] both *in vitro* and *in vivo*.

As this article is rather lengthy, readers not interested in technical details may prefer to move straight to the Results section. An appendix contains useful formulae for obtaining timescales for the dominant nucleation collapse process for small proteins.

Description of the model

The model used in our simulations is a variant of the one introduced in our previous studies [16,17]. We use continuum minimal model representation of a polypeptide chain. In these classes of models, only the principle features of proteins responsible for imparting stability are retained. These include hydrophobic forces, excluded volume interactions, bond angle and dihedral angle degrees of freedom. The simplified model can be thought of as a coarse-grained representation containing only the α-carbons of the protein molecule. The polypeptide is modeled as a chain consisting of N connected beads with each corresponding to a set of particular α-carbons in a real protein. In order to simplify the force field, we assume that sequence is essentially built from residues of three types: hydrophobic (B), hydrophilic (L), and neutral (N). Our previous studies have established that this three-letter code can be used to construct the basic structural motifs in proteins, namely α-helix and β-turn [16,32,34]. In this study, we mimic the diversity in the hydrophobic residues in proteins using a dispersion in the interactions between B residues (see below).

The potential energy of a conformation, which is specified by the set of vectors $\{\vec{r}_i\}$, i=1,2...N, is taken to be of the following form:

$$E_p(\{\vec{r}_i\}) = V_{BL} + V_{BA} + V_{DIH} + V_{NON} \tag{2}$$

where V_{BL}, V_{BA}, V_{DIH}, and V_{NON} correspond to bond length potential, bond angle potential, dihedral angle potential, and nonbonded potential, respectively. A brief summary of these interactions is given below.

Bond length potential

In our previous studies, we assumed the length of the covalent bond connecting the successive beads to be fixed. The constraint of fixed bond length, which was enforced using the RATTLE algorithm [35], proves to be computationally demanding. In the present study, we use a stiff harmonic potential between successive residues, which keeps the bond length approximately fixed, i.e.:

$$V_{BL} = \sum_{i=1}^{N-1} \frac{k_r}{2}(|\vec{r}_{i+1} - \vec{r}_i| - a)^2 \tag{3}$$

where $k_r = 100\epsilon_h/a^2$, a is the average bond length between two beads, and ϵ_h, the average strength of the hydrophobic interaction, is the unit of energy in our model. We have verified that using the potential in eq. 3 gives the same results for the sequence that has been previously studied [16].

Bond angle potential

The potential for the bending degrees of freedom, describing the angle between three successive beads i, i+1, i+2, is taken to be:

$$V_{BA} = \sum_{i=1}^{N-2} \frac{k_\theta}{2}(\theta_i - \theta_0)^2 \tag{4}$$

where $k_\theta = 20\epsilon_h/(rad)^2$ and $\theta_0 = 1.8326\ rad$ or 105°.

Dihedral angle potential

This potential describes the ease of rotation around the angle formed between four consequent beads. This degree of freedom is largely responsible in determining secondary structures in a polypeptide chain [36]. The i^{th} dihedral angle ϕ_i is formed between vectors $\vec{n}_i = (\vec{r}_{i+1,i} \times \vec{r}_{i+1,i+2})$ and $\vec{n}_{i+1} = (\vec{r}_{i+2,i+1} \times \vec{r}_{i+2,i+3})$, i.e. it is the angle between the plane defined by beads i, i+1, i+2 and the one spanned by beads i+1, i+2, i+3. The vector $\vec{r}_{i,i+1} = \vec{r}_{i+1} - \vec{r}_i$. The general form of the potential describing the dihedral angle degrees of freedom is well known [37] and can be represented as:

$$V_{DIH} = \sum_{i=1}^{N-3} \left[A_i(1 + \cos\phi_i) + B_i(1 + \cos 3\phi_i) \right] \tag{5}$$

If two or more of the four beads in defining ϕ_i are neutral (N) the A_i and B_i are taken to be $0\epsilon_h$ and $0.2\epsilon_h$, respectively. For all other cases, $A_i = B_i = 1.2\epsilon_h$. For the larger values of A_i and B_i, the *trans* state is preferred, and this leads to the formation of extended conformation. The presence of neutral residues, which are introduced so that loop formation is facilitated, has the effect of decreasing the barrier and energetic differences between the *trans* and *gauche* states [34].

Nonbonded potential

The nonbonded potentials arise between pairs of residues that are not covalently bonded. These forces together with those arising from the dihedral angle degrees of freedom (which provide favorable local interactions for the formation of secondary structures) are responsible for the overall formation of the three-dimensional topology of the polypeptide chain.

We take simple forms to represent the nonbonded interaction terms. We assume that the effective potential describing the interaction between the residues i and j ($|i - j| \geq 3$)

depends on the type of residues involved. The total non-bonded potential is written as:

$$V_{NON} = \sum_{i=1}^{N-3} \sum_{j=i+3}^{N} V_{ij}(r) \tag{6}$$

where $r = |\vec{r_i} - \vec{r_j}|$. The potential between two L beads or between a (L,B) pair is taken to be:

$$V_{L\alpha}(r) = 4\epsilon_L \left[\left(\frac{a}{r}\right)^{12} + \left(\frac{a}{r}\right)^{6} \right] (\alpha = L \text{ or } B) \tag{7}$$

where $\epsilon_L = 2/3\epsilon_h$. This potential is purely repulsive with a value of $2\epsilon_h$ at $r = 2^{1/6}a$, which is the location of the minimum in the hydrophobic potential (see eq. 9). The presence of the r^{-6} term gives rise to a potential that is longer ranged than the usual r^{-12} term. The additional term may be interpreted to arise from the hydration shells around the hydrophilic residues.

The interaction between the neutral residues and the others is expressed as:

$$V_{N\alpha}(r) = 4\epsilon_h \left(\frac{a}{r}\right)^{12} (\alpha = N,\ L,\ \text{or } B) \tag{8}$$

If both the residues are hydrophobic (B) the potential of interaction is taken to be:

$$V_{BB}(r) = 4\lambda\epsilon_h \left[\left(\frac{a}{r}\right)^{12} - \left(\frac{a}{r}\right)^{6} \right] \tag{9}$$

where ϵ_h determines the strength of the hydrophobic interaction. The above form for $V_{BB}(r)$ can be thought of as approximate representation (capturing the primary minimum) of the potential of mean force between spherical hydrophobic spheres in water [38].

The dimensionless parameter λ is assumed to have a Gaussian distribution:

$$P(\lambda) = \frac{1}{(2\pi\Lambda^2)^{1/2}} \exp\left(-\frac{(\lambda - \lambda_0)^2}{2\Lambda^2} \right) \tag{10}$$

The mean value of $\lambda_0 = 1$. The introduction of the distribution in the strength of the hydrophobic interaction creates diversity among hydrophobic species and hence provides a better caricature of proteins. The standard deviation Λ controls the degree of heterogeneity of the hydrophobic interactions and if its value becomes too large then the unambiguous division of residues into three distinct types becomes difficult. Consequently, we keep the value of Λ small enough so that the prefactor $4\lambda\epsilon_h$ is in general greater than $4\epsilon_L$ (see eqs 7,9). This also assumes that the interaction between hydrophobic residues remains attractive at the separation corresponding to Lennard–Jones minimum. For large values of Λ (not used in our present study), the distribution function in eq. 10 has to be truncated at some positive value of λ so that the $\lambda\epsilon_h$ does not become negative.

In our earlier studies [16,17,34], we used $\Lambda = 0$ and hence all B beads were identical. Thus, our previous studies correspond exactly to a three-letter code. The current potential function with random hydrophobic interaction gives more specificity to the interactions and yet preserves the overall hydrophobic interactions as the driving force for structure formation.

Simulation methods

Langevin dynamics

Following our earlier work, we have used Langevin dynamics for simulating folding kinetics [16,34]. We include a damping term in the equation of motion with a properly chosen friction coefficient ζ and the Gaussian random force to balance the energy dissipation caused by friction. The equation of motion written for the generalized coordinate x is given by:

$$m\ddot{x} = -\zeta\dot{x} + F_c + \Gamma \equiv F \tag{11}$$

where $F_c = -\partial E_p/\partial x$ is the conformation force, which is a negative gradient of potential energy with respect to the coordinate x, Γ is the random force having a white noise spectrum, and m is the mass of a bead. The equation of motion (eq. 11) is numerically integrated using the velocity form of the Verlet algorithm [39]. If the integration step is h, the position of a bead at the time $t+h$ is expressed through the second order in h as:

$$x(t+h) = x(t) + h\dot{x}(t) + \frac{h^2}{2m}F(t) \tag{12}$$

Similarly, the velocity $\dot{x}(t+h)$ at the time $t+h$ is given by:

$$\dot{x}(t+h) = \left(1 - \frac{h\zeta}{2m}\right)\left(1 - \frac{h\zeta}{2m} + \left(\frac{h\zeta}{2m}\right)^2\right)\dot{x}(t) + \frac{h}{2m}\left(1 - \frac{h\zeta}{2m} + \left(\frac{h\zeta}{2m}\right)^2\right) \times (F_c(t) + \Gamma(t) + F_c(t+h) + \Gamma(t+h)) + o(h^2) \tag{13}$$

Because we assume that the random force Γ has a white noise spectrum, the autocorrelation function $\langle\Gamma(t)\Gamma(t')\rangle$ is expressed in the form:

$$\langle\Gamma(t)\Gamma(t')\rangle = 2\zeta k_B T\delta(t - t') \tag{14}$$

Since the equation of motion (eq. 11) is discretized and solved numerically, this formula can be rewritten as:

$$\langle\Gamma(t)\Gamma(t + nh)\rangle = \frac{2\zeta k_B T}{h}\delta_{0,n} \tag{15}$$

where $\delta_{0,n}$ is the Kronecker delta and n = 0,1,2.... Thus, in the context of this model, changing the temperature of the system essentially means changing the standard variance in the Gaussian distribution of the random force Γ.

Temperature is measured in the units of ϵ_h/k_B. In the underdamped limit, i.e. when $\zeta\dot{x}$ is negligible compared to the inertial term in the equation of motion (eq. 11), a natural choice of the unit of time is $\tau_L = (ma^2/\epsilon_h)^{1/2}$. The simulations have been done in low to moderate friction limit which, in the rate theory of reactions, would correspond to the energy diffusion regime. The integration step used in the equation of motion is taken to be $h = 0.005\tau_L$. All the sequences were studied at two values of the friction coefficient $\zeta_L = 0.05m\tau_L^{-1}$ and $\zeta_M = 100\zeta_L = 5m\tau_L^{-1}$. The relation between the time unit τ_L and the folding timescales in real proteins as well as the range of ζ used in this study are discussed in the Appendix. The Appendix also gives estimates for certain timescales in the folding kinetics of proteins using the simulation results and theoretical arguments. In our simulations, the mass of residue m, the bond length a, the hydrophobic energy constant ϵ_h, and the Boltzmann constant k_B are set to unity.

Determination of native conformation

For each sequence, we determined the native conformation by adapting the procedure similar to that used in our recent work on lattice model of proteins [23]. (The database of sequences generated is described below.) As in our earlier works [16,34], we have used a combination of slow cooling and simulated annealing to determine the native conformation. The chain is initially heated to T = 5.0 and equilibrated at this temperature for $2000\tau_L$. The temperature is then quenched to T = 1.0 and the chain is reequilibrated for an additional $2000\tau_L$. This process of quenching the chain from T = 5.0 to T = 1.0 was repeated several times so that we generated a set of independent conformations at T = 1.0. These structures are used as starting conformations for the slow cooling process. In order to ascertain that the starting conformations are independent, the overlap between a pair of these conformations (see eq. 17) averaged over all distinct pairs denoted by $\overline{\chi}$ is calculated. This yields $\overline{\chi} \sim 0.9$ that roughly corresponds to the value of χ for a pair of randomly generated conformations. The temperature of the system in the simulations, starting from one of the well equilibrated conformations at T = 1.0, is slowly decreased to T = 0.0. In the process of reaching T = 0.0, the energies of the conformations are recorded. This process is repeated for several (typically 10) initial conformations. The conformation with the lowest energy is assumed to be the native state for the sequence. After determining the native conformation by this method, we raised the temperature from 0.0 to 0.2 in $1000\tau_L$ and then lowered it to T = 0.0, i.e. we performed a simple simulated annealing procedure. In all instances, the resulting structure and the energy coincided with those obtained by the slow cooling protocol. It should be emphasized that this method cannot guarantee that the structures are indeed global energy minima. However, the determination of native structures for these sequences by other optimization techniques leads to the same structures [40]. Thus, we are fairly certain that the structures found by this method indeed are the lowest energy structures for our model.

Thermodynamic properties

We and others have shown that each foldable sequence is characterized by two natural temperatures [1,4,7,16,24]. One of them is T_θ, below which the chain adopts more or less compact conformation. The transition at T_θ is (usually) second order in character suitably modified by finite size effects. Following our earlier studies, T_θ is located by determining the temperature dependence of the heat capacity [16,23,24]:

$$C_v = \frac{\left\langle E^2 \right\rangle - \left\langle E \right\rangle^2}{T^2} \tag{16}$$

The location of the peak in C_v is taken to be T_θ. Previous studies have shown that at $T \approx T_\theta$ the radius of gyration changes dramatically, reaching a value roughly coinciding with that for a compact conformation [16,34]. This, of course, is usually taken to be a signature of 'collapse' transition in homopolymers [41].

The second crucial temperature is the folding transition temperature, T_F. There are several ways of calculating T_F, all of which seem to give roughly similar estimates [7,24,42]. We use the fluctuations in the structural overlap function to estimate T_F. The structural overlap function is defined as:

$$\chi = 1 - \frac{2}{N^2 - 5N + 6} \sum_{i=1}^{N-3} \sum_{j=i+3}^{N} \Theta\left(\epsilon - \left| r_{ij} - r_{ij}^N \right|\right) \tag{17}$$

where r_{ij} is the distance between the beads i and j for a given conformation, r_{ij}^N is the corresponding distance in the native conformation, and $\Theta(x)$ is the Heavyside function. If $|r_{ij} - r_{ij}^N| \leq \epsilon$ then the beads i and j are assumed to form a native contact. In our simulations, we take $\epsilon = 0.2a$.

It follows from the definition of χ that at finite temperatures $\langle\chi(T)\rangle$, the thermal average, is in general non-zero. The folding transition temperature is obtained from the temperature dependence of the fluctuations in χ:

$$\Delta\chi = \left\langle \chi^2(T) \right\rangle - \left\langle \chi(T) \right\rangle^2 \tag{18}$$

For sequences with a unique ground state, $\Delta\chi$ exhibits a peak at $T \simeq T_F$ [24]. It has been shown that for these simple off-lattice models, this transition is a finite size first-order phase transition [16]. Our previous lattice model studies have shown that T_F obtained from the tem-

perature dependence of $\Delta\chi$ is in general slightly smaller than that calculated from the midpoint of $\langle\chi(T)\rangle$ or other suitable order parameters [43].

The thermodynamic properties such as $\langle\chi(T)\rangle$ and total energy $\langle E(T)\rangle$ (the sum of kinetic and potential energies) are calculated using time averages over sufficiently long trajectories. The trajectories which are generated in the search for the native structure can be used to get an approximate estimate of the temperature interval T_l,T_h, which includes the temperatures T_θ and T_F. In all cases, we set $T_h = 1.0$, while T_l varies from 0.3 to 0.4. Each trajectory starts with the same zigzag initial conformation. The chain is then heated at $T = 5.0$ and brought to equilibrium at $T = 1.0$. The method of slow cooling employed for calculating thermodynamic averages is identical to that presented in [43]. The system is periodically cooled (starting at the temperature T_h) by an amount ΔT. The time τ_{max} is the time of running the simulations at the fixed temperature, $T_i = T_h - i\Delta T$, where $i = 0,1,2....$ In this study, we have set $\Delta T = 0.02$, the time $\tau_{max} = 2500\tau_L$, and the equilibration time after the change of the temperature by ΔT to be $\tau_{eq} = 250\tau_L$. These values are used with most sequences within the entire temperature interval (T_l,T_h). The thermodynamic values for one particular initial condition i are calculated as:

$$\overline{f}_i(T) = \frac{1}{\tau_{av}} \int_{\tau_{eq}}^{\tau_{eq}+\tau_{av}} f_i(T,t)dt \qquad (19)$$

where $\tau_{av} = \tau_{max} - \tau_{eq}$. The equilibrium thermodynamic value is obtained by averaging over a number of initial conditions:

$$\langle f(T)\rangle = \frac{1}{M}\sum_{i=1}^{M} \overline{f}_i(T) \qquad (20)$$

We found that $M = 50$ was sufficient to obtain accurate results for equilibrium properties. For most sequences, the values of parameters (see above) used in the course of equilibration were large enough to obtain converged results. In some instances, the equilibration times had to be increased to obtain converged results. The functions $\langle E(T)\rangle$, $C_v(T)$, and $\Delta\chi(T)$ are obtained by fitting the data by polynomials, and $\langle\chi(T)\rangle$ was fit with two hyperbolic tangents.

We should note that due to the intrinsic heterogeneity of these systems non-ergodicity effects often manifest themselves [16]. If this is the case, we need to do weighted averaging of thermodynamic quantities, as described elsewhere [16,44], to get converged results. This issue was not encountered for the sequences that were examined in this study.

It is obvious that the thermodynamic quantities are independent of the underlying dynamics provided the dynamics yields the Boltzmann distribution at $t \to \infty$. Since the thermodynamics are determined only by the Boltzmann factor $\exp(-E_p/k_BT)$, it is convenient to determine them at low friction, where the sampling of conformation space appears to be more efficient [16,34].

Database of sequences

For our model, one can, in principle, generate an infinite number of sequences because of the continuous distribution of the effective hydrophobic interactions (see eq. 10). The vast majority of such sequences would be random and hence would not fold to a unique native state on finite timescale. Our goal is to obtain a number of these sequences with the characteristic temperatures T_θ and T_F such that they span a reasonable range of σ (see eq. 1). It is clear that merely creating random sequences will not achieve this objective. In general, random sequences would take extraordinarily long times to fold. It is known that foldable sequences (those that reach the native state in finite times) are designed to have a relatively smooth energy landscape. Such sequences may, in fact, be minimally frustrated [45] or have compatible long-range and short-range interactions [46]. Thus, in order to generate sequences that span the range of σ and which are foldable, we used the most primitive design procedure in the inverse protein folding problem [47,48]. Because our objective is not to provide the most optimal solution to the inverse folding problem, the more reliable methods introduced recently were not utilized [49].

In all our studies, the number of beads $N = 22$. The composition of all sequences is identical, i.e. all of them contain 14 hydrophobic beads, 5 hydrophilic beads, and 3 neutral beads. The sequences in this model differ from each other because of the precise way in which these beads are connected. In addition, due to the distribution of hydrophobic interactions, not all hydrophobic beads are identical. The latter condition also introduces diversity among sequences.

The method for creating the database of sequences is as follows. The first sequence, A $[B_9N_3(LB)_5]$, has already been studied in our previous work [16]. This allows us to ascertain that our modified model (incorporating stiff harmonic bond length potential instead of the RATTLE algorithm) yields results consistent with our earlier studies. This sequence has zero Λ, so all hydrophobic residues are identical. All other sequences (to be used as starting conditions for the Monte Carlo optimization procedure) were generated at random with different standard deviations Λ, but preserving the same composition, i.e. 14 B, 5 L, and 3 N beads. By 'random generation' of a sequence we mean that a sequence is randomly constructed from the beads of three types, and the values of the parameter λ specifying nonbonded interactions between hydrophobic residues in a sequence (eq. 9) were obtained using Gaussian distribution (eq. 10). Specifically, we used $\Lambda = 0$ (one sequence), $\Lambda = 0.1$ (one sequence), $\Lambda = 0.17$ (one sequence), and $\Lambda = 0.3$ (five sequences).

The next step in creating the database of sequences is the choice of the target conformation, the precise choice of which is rather arbitrary because the only natural requirement is that it should be compact. However, due to intrinsic propensity toward bend formation near clusters of N residues, it seems reasonable to restrict the choice of target conformation topology to that with a single U-turn. Obviously, the number of residues in a sequence $N = 22$ allows us to define much more complicated topologies featuring multiple U-turns. In order to avoid the comparison of folding behavior of the sequences with the native conformations of different topologies, the only criterion for selecting target conformations is that they must have a single U-turn and be reasonably compact. The role of the topology of the ground state in determining the folding kinetics will be addressed in a future paper. Thus, using the conditions described above, we have selected the target conformations from the database of low energy structures found in the course of slow cooling simulations.

Once a target conformation is chosen, the Monte Carlo algorithm in sequence space [47,48] is used to obtain an optimal sequence by means of the primitive inverse design procedure, i.e. the sequence that has the lowest energy and is compatible with the target conformation. There is no guarantee that this procedure is by any means the best way of designing optimal sequences, as has been pointed out by Deutsch and Kurosky [49], but for our purposes this naive procedure suffices. The main idea is to perform small random permutations of a sequence while keeping its composition fixed and accepting (or rejecting) new sequences with respect to the Boltzmann factor $P = \exp(-\Delta E/k_B T)$, where ΔE is the energy variation due to permutation and T is the temperature of Monte Carlo optimization scheme [47,48]. Hence, this algorithm is aimed at lowering the energy of the target conformation. The sequence providing the lowest energy at the target conformation is chosen as the desired sequence and used for further analysis. The control parameter which specifies the degree (or 'quality') of optimization is the temperature T. For most sequences we have used low value of $T = 0.2$. We have made several attempts to run the Monte Carlo algorithm in a sequence space by gradually decreasing the temperature from high values of $T \gtrsim 1.0$ up to the very low values $T \lesssim 0.01$. This method, however, does not often provide lower energies of target conformations than that based on quenching the temperature at a certain value, and this is probably due to moderate ruggedness of the energy landscape in sequence space. It must be emphasized that the optimization scheme does not guarantee that the target conformation is actually the native state of the optimized sequence. This should be checked in the course of molecular dynamics simulations (see below).

The nine sequences (A–I) obtained using the procedure described above are listed in Table 1. Of these, eight (B–I) were generated using the Monte Carlo method in sequence space. All sequences have different native conformations, except the pairs of sequences A,D and B,C, which share the same native state. It must also be emphasized that although sequences F–H have identical distribution of beads, they differ from each other with respect to the strength of hydrophobic interactions, i.e. they have different sets of prefactors λ (eq. 10).

Simulation temperature

In order to compare the rates of folding for different sequences, it is desirable to subject them to identical folding conditions. The equilibrium value of $\langle\chi(T)\rangle$ measures the extent to which the conformation at a given temperature T is similar to the native state. At sufficiently low temperature, $\langle\chi(T)\rangle$ would approach zero, but the folding time may be far too long. We chose to run our folding simulations at a sequence-dependent simulation temperature T_s which is subject to two conditions: T_s be less than T_F for a specified sequence so that the native conformation has the highest occupation probability, and the value of $\langle\chi(T=T_s)\rangle$ be a constant for all sequences, i.e.:

$$\langle\chi(T = T_s)\rangle = \alpha \tag{21}$$

In our simulations, we chose $\alpha = 0.26$ and for all the sequences studied $T_s/T_F < 1$. This general procedure for selecting the simulation temperatures has already been used in recent studies of folding kinetics using lattice models [42,43,50].

An alternative way of choosing T_s is to assume that the probability of occupation of the native state be the same for all sequences. In our previous study [43] on lattice models, we used this method for a small number of sequences. The trends in the folding times at the resulting simulation temperatures (as well as the kinetics) were very similar to those found when eq. 21 is used to determine T_s.

It is also possible to keep the simulation temperature constant for all sequences. Because T_F and T_θ vary greatly depending on sequence, such a choice would not ensure that the probability of being in the native conformation is roughly the same for all sequences or that all sequences are qualitatively similar to the same extent. In other words, the folding conditions for the sequences would effectively be different if the temperature is held constant. This argument also implies that the statistical trends of folding kinetics with respect to intrinsic sequence-dependent properties are expected to hold over only an optimal range of folding conditions, which in simulations are entirely determined by temperature.

Monitoring folding kinetics

The simulation procedure for obtaining folding kinetics resembles the slow cooling method apart from the one principal difference that after heating the chain it is quenched

Table 1

Sequences and their parameters studied in simulations.

Label	Sequence	Λ	T_F	T_θ	σ
A	$B_9N_3(LB)_5$	0	0.20	0.58	0.65
B	$B_8(NL)_2NBLB_3(LB)_2$	0	0.30	0.62	0.51
C	$B_8NLN_2LBLB_3(LB)_2$	0.3	0.38	0.72	0.47
D	$B_9N_2LNB(LB)_4$	0.17	0.36	0.62	0.41
E	$LB_7NBNLN(BL)_2B_3LB$	0.1	0.40	0.66	0.39
F	$LB_9(NL)_2NBLB_3LB$	0.3	0.46	0.76	0.39
G	$LB_9(NL)_2NBLB_3LB$	0.3	0.62	0.78	0.20
H	$LB_9(NL)_2NBLB_3LB$	0.3	0.59	0.80	0.26
I	$LBNB_3LB_3N_2B_2LBLB_3LB$	0.3	0.54	0.62	0.14

B, hydrophobic; *L*, hydrophilic; *N*, neutral.

to the temperature T_s, defined from the condition $\langle\chi(T{=}T_s)\rangle = 0.26$. The temperature is held constant after the quench to $T = T_s$. The duration of the folding simulations depends on the rate of folding of a particular sequence and is typically on the order of $10^4\tau_L$. For each sequence we generated between 100 and 300 independent trajectories. The folding kinetics is monitored using the fraction of trajectories $P_u(t)$ that does not reach the native conformation at time t:

$$P_u(t) = 1 - \int_0^t P_{fp}(s)ds \tag{22}$$

where $P_{fp}(s)$ is the distribution of first passage times:

$$P_{fp}(s) = \frac{1}{M}\sum_{i=1}^{M}\delta(s-\tau_{1i}) \tag{23}$$

τ_{1i} denotes the first passage time for the i^{th} trajectory, i.e. the time when a sequence adopts native state for the first time. It is easy to show that the mean first passage time τ_{MFPT} to the native conformation (which is roughly the folding time, τ_F) can be calculated as:

$$\tau_{MFPT} = \int_0^\infty tP_{fp}(t)dt = \int_0^\infty P_u(t)dt \tag{24}$$

The mean first passage time τ_{MFPT} can also be calculated using τ_{1i} for the M trajectories so that:

$$\tau_{MFPT} = \frac{1}{M}\sum_{i=1}^{M}\tau_{1i} \tag{25}$$

We find that for all the sequences, $P_u(t)$ can be adequately fit with several exponentials of the form:

$$P_u(t) = \Phi\exp\left(-\frac{t}{\tau_{FAST}}\right) + \sum_k a_k\exp\left(-\frac{t}{\tau_k}\right) \tag{26}$$

where the sum is over the dominant misfolded structures, τ_k are the timescales for activated transition from one of the misfolded structures to the native state, and $\Sigma_k a_k = 1 - \Phi$. In this study, we report τ_{MFPT} using eq. 24 by fitting $P_u(t)$ according to eq. 26. We have explicitly verified that eq. 24 (with $P_u(t)$ given by eq. 26) and eq. 25 yield practically identical results.

It has been shown in a series of articles that multiexponential fit of the function $P_u(t)$ can be understood in terms of the KPM [8,16,17,20] and is indicative of the distribution of timescales in the refolding of biomolecules. According to KPM, a fraction of molecules Φ fold to the native conformation very rapidly, while the remainder (1 – Φ) approach the native state via a complex three-stage multipathway mechanism (TSMM). Therefore, the time constants τ_{FAST} and τ_k can be interpreted as the characteristic folding times of the fast and slow phases, respectively. When appropriate fit of the function $P_u(t)$ with exponentials is performed, the calculation of the mean first passage time τ_{MFPT} becomes straightforward.

An alternative way to calculate folding times is based on the analysis of the time dependence of the overlap function. The overlap function χ is constantly calculated during simulations and its average time dependence is obtained as:

$$\langle\chi(t)\rangle = \frac{1}{M}\sum_{i=1}^{M}\chi_i(t) \tag{27}$$

where $\chi_i(t)$ is the value of χ for the i^{th} trajectory at time t. We find that $\langle\chi(t)\rangle$ can also be fit by a sum of exponentials (usually by one or two, see [23,24]):

$$\langle\chi(t)\rangle = a_1\exp\left(-\frac{t}{\tau_1}\right) + a_2\exp\left(-\frac{t}{\tau_2}\right) \tag{28}$$

Here, τ_1 gives the estimate for the timescale of the NCNC process [16,17], while the largest time constant in the fit τ_2 serves as an estimate of the folding time τ_F. For most sequences, biexponential fit provides the most accurate results. However, several sequences (typically ones with relatively small values of σ) demonstrate a clear single exponential behavior of $\langle\chi(t)\rangle$. In some cases, there are additional slow components present on larger timescales as well. It has been shown in our previous papers that defining folding times using the functions $P_u(t)$ or $\langle\chi(t)\rangle$ yields qualitatively similar results [16,43]. The same conclusion is valid for this model as well. Thus, τ_{FAST} and the largest τ_k in eq. 26 are roughly proportional to τ_1 and τ_2, respectively.

Results

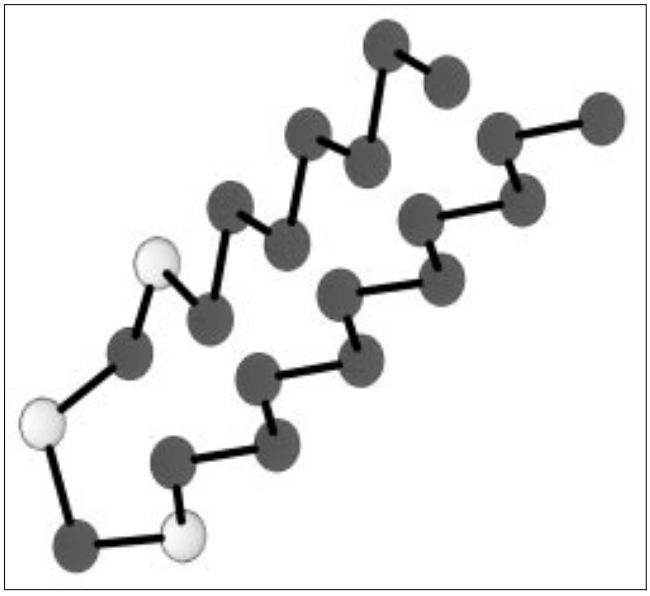
Thermodynamic properties

In this section, we present the results on thermodynamics and kinetics of folding. Using the methodology described

above, we studied nine sequences, eight of which were generated by performing Monte Carlo simulations in sequence space. The native conformation of each sequence was determined, and the example of the native conformation for sequence G is given in Figure 1. This demonstrates that all three neutral residues (shown in grey) are concentrated in the turn region. It is also clearly seen that hydrophobic residues (shown in blue) tend to be in close contact to each other due to their inherent attractive interaction, while hydrophilic residues (shown in red) point outwards.

For each sequence, we calculated the two characteristic equilibrium temperatures, collapse transition temperature T_θ and folding transition temperature T_F from the temperature dependence of C_v and $\Delta\chi$, respectively. The plots of $\langle\chi\rangle$ and $\langle E\rangle$ were also obtained. Figure 2 displays these functions for sequence G. The plot of $\langle\chi(T)\rangle$ (Fig. 2a) indicates that at high temperatures $0.8 \lessapprox T \lessapprox 1.0$, the value of $\langle\chi\rangle \approx 0.8$, so the chain has a negligible amount of the native structure. It had already been shown [34] that under these conditions the polypeptide chain is in a random coil conformation. The overlap function gradually decreases with temperature and at $T = 0.3$ it reaches a value below 0.2. We do not plot $\langle\chi(T)\rangle$ for $T < 0.3$, because at such low temperatures it is difficult to obtain reliable thermodynamic averages due to non-ergodicity problems. Fortunately, this is not necessary for our study because all characteristic temperatures T_θ, T_F and the simulation temperature T_s are relatively high. The peak of the specific heat C_v (Fig. 2d) which corresponds to the collapse transition temperature T_θ is at 0.78. At this temperature, the protein undergoes a transition from an extended coil state to compact conformation. In fact, we calculated the radius of gyration $\langle R^2g\rangle$ as a function of temperature for a few sequences and found that at $T \approx T_\theta$ it shows a sudden drop in accord with earlier and more recent studies [16,34]. However, at T_θ the overlap function is still relatively large $\langle\chi\rangle \approx 0.7$. The fluctuation of the overlap function $\Delta\chi$ achieves a maximum at $T = 0.62$ and this is taken to be the folding temperature T_F. The value of T_F calculated from the midpoint of $\langle\chi\rangle$ (i.e. when $\langle\chi\rangle$ is about 0.5) is also around 0.62. In general, we have found that T_F obtained from the peak of $\Delta\chi$ is slightly lower than that calculated from the temperature dependence of similar measures such as $\langle\chi\rangle$ [43]. It was demonstrated that this temperature corresponds to first-order folding transition to the native conformation [16,24]. By monitoring $\langle\chi(t)\rangle$ for several individual trajectories under equilibrium conditions at T_F we find that the protein fluctuates between the native and disordered conformations. All nine sequences show similar behavior from which the various thermodynamic parameters can be easily extracted. The parameter σ (see eq. 1) for the nine sequences ranges from 0.14 to 0.65. Thus, a meaningful correlation between the folding time and σ, which is one of the major aims of this study, can be established.

Figure 1



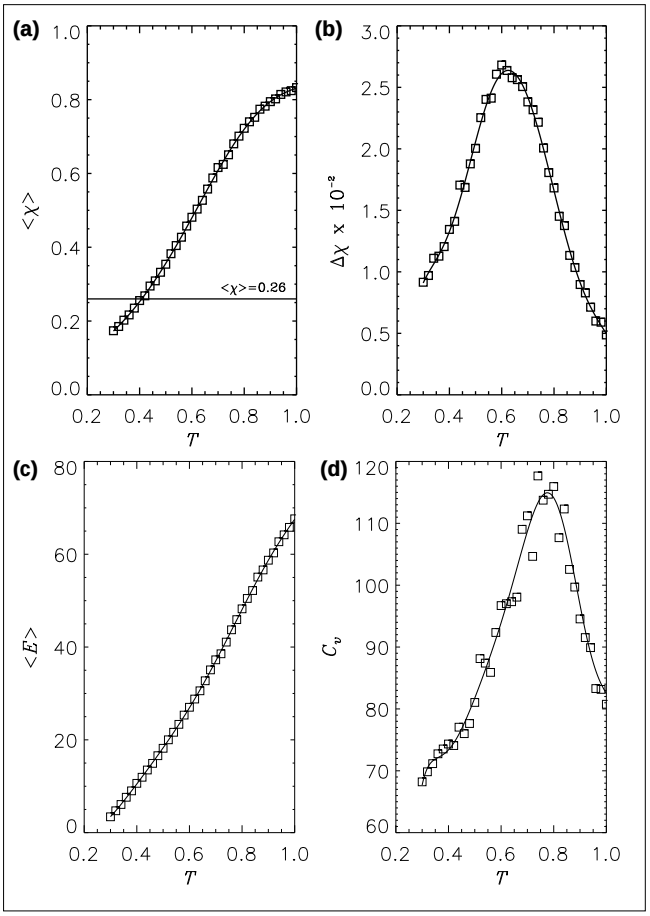
The β-type native structure of sequence G: $LB_9(NL)_2NBLB_3LB$. *B*, hydrophobic (blue); *L*, hydrophilic (red); *N*, neutral (grey). In the turn region, the chain backbone adopts *gauche* conformations.

The simulation temperature T_s for the sequence G defined by eq. 21 is found to be 0.41. In Figure 3, we present the dependence of T_s on the parameter σ. It is seen that the simulation temperature T_s is a decreasing function of σ. Thus, high values of T_s are found for sequences with small values of σ, and this prompts us to anticipate that such sequences are fast folders (see below). However, it was argued [43] that this correlation must be viewed as statistical. This implies that if two sequences have close values of σ, then a precise correlation with T_s is not always expected. On the other hand, if a large number of sequences spanning a range of σ are generated, then we expect a statistical correlation to hold. We also expect these conclusions to hold over a range of temperatures that are favorable for folding. The present off-lattice studies and those based on lattice models [23,43] confirm this expectation.

Dependence of T_θ and T_F on sequence

One of the major results in this study is that the folding times for all sequences correlate extremely well with σ (cf. eq. 1). Therefore, it is of interest to investigate how T_θ and T_F vary with the sequence. It seems reasonable to assert that the folding temperature T_F depends rather sensitively on the precise sequence. In fact, it has been argued that to a reasonable approximation, T_F is determined by the nature of the low energy spectrum (a sequence-dependent property), at least in lattice models [51]. The sensitive dependence of T_F on the sequence is explicitly confirmed in this paper and in the previous lattice models [43,51]. In

Figure 2

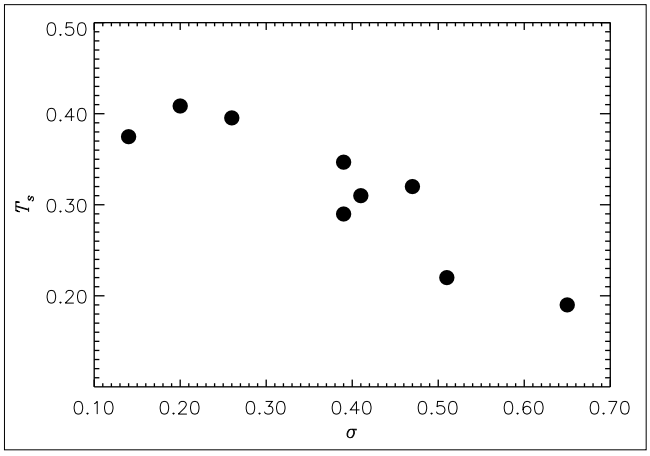


The temperature dependence of the thermodynamic quantities for sequence G calculated using a slow cooling method. (a) Overlap function $\langle\chi(T)\rangle$. (b) Fluctuations of the overlap function $\Delta\chi(T)$. (c) Energy $\langle E(T)\rangle$. (d) Specific heat $C_v(T)$. The peaks in the graphs of $\Delta\chi(T)$ and $C_v(T)$ correspond to the folding transition and collapse transition temperatures, T_F and T_θ, respectively.

Table 1, we display T_F for the nine sequences. The values of T_F range from 0.20 to 0.62. Thus, the largest T_F is about three times larger than the smallest value.

It might be tempting to think that T_θ should be insensitive to the sequence and should essentially be determined by the composition of the sequence. This expectation arises especially from heteropolymer theory [52] (which essentially ignores short length scale details), according to which T_θ is determined by the average excluded volume interactions, v_0, and the average strength of hydrophobic interactions, $\lambda_0\epsilon_h$. Both these values are expected to be roughly constant, especially if the sequence composition is fixed. The determination of T_θ for a polypeptide chain based on these arguments ignores surface terms and may, in fact, be valid in the thermodynamic limit, i.e. when the number of beads tends to infinity. However, polypeptide chains are finite in size and hence the nature of surface

Figure 3



The dependence of the simulation temperature T_s, as defined by eq. 21, on the parameter $\sigma = (T_\theta - T_F)/T_\theta$.

residues, which depends on the precise sequence, is critical in the determination of T_θ. This is borne out in our simulations. In Table 1, we also display T_θ for the nine sequences. Although the largest T_θ is only approximately 1.5 times (as opposed to a factor of three for T_F) larger than the lowest, it is clear that T_θ is very sequence dependent even though the composition for all sequence is identical. All the sequences have 14 hydrophobic residues. Both T_θ and T_F are determined not only by the intrinsic sequence but also by external conditions. In fact, T_θ and T_F, and consequently σ, can be manipulated by altering the external solvent conditions (pH, salt, etc.). It therefore follows that a single foldable sequence can have very different values of σ depending on the solvent conditions and hence can exhibit very different kinetics.

It is interesting to obtain estimates for T_θ and T_F using realistic values of ϵ_h, the average strength of hydrophobic interaction. From Table 1 we note that the range of T_θ is $0.58–0.80\epsilon_h/k_B$ with the lower values corresponding to sequences with larger σ. The value of ϵ_h ranges from 1 to 2 kcal mol^{-1}. Assuming that $\epsilon_h \approx 2$ kcal mol^{-1}, the range of T_θ is 48–67°C. It appears that the better designed sequences (ones with smaller σ values) have more realistic values of T_θ. Similarly, the range of T_F for better designed sequences is 33–50°C. These estimates suggest that optimized sequences can fold over a moderate range of temperatures rapidly and with relatively large yield. These expectations are explicitly demonstrated here.

Kinetics of folding: the kinetic partitioning mechanism

We studied the folding kinetics using the function $P_u(t)$, which gives the fraction of unfolded molecules (trajectories) at time t. We also computed the time dependence of $\langle\chi(t)\rangle$ to gain additional kinetic information concerning the

approach to the native conformation. The function $P_u(t)$ has been obtained for each sequence at the simulation temperature T_s from the analysis of a large number of individual trajectories (M = 100–300) starting with different initial conditions. The resulting plots of $P_u(t)$ were fitted with a sum of exponentials (one, two, or three) and the mean first passage time τ_{MFPT} (taken to be equal to τ_F) for each sequence was calculated. In general, it is found that after a short transient time $P_u(t)$ is extremely well fit by a sum of exponentials (cf. eq. 26). The partition factor, Φ, gives the fraction of molecules that reaches the native conformation on the timescale τ_{FAST} by a NCNC mechanism, τ_k (>> τ_{FAST}) being the timescale over which the remaining fraction $1 - \Phi$ reaches the native state [16,17,43].

Based on fairly general theoretical considerations, it has been shown that σ [= $(T_\theta - T_F)/T_\theta$] can be used to discriminate between fast and slow folding sequences [21,24]. This has been confirmed numerically for lattice models [23,43,51]. We classify fast-folding sequences as those with relatively large values of Φ ($\gtrsim$ 0.9). These sequences reach the native conformation without forming any discernible intermediates and essentially display a two-state kinetic behavior. The plot of $P_u(t)$ for one of the sequences (sequence G), which can be fit with only one exponential in eq. 26, is presented in Figure 4a. It is obvious that Φ depends on the sequence (via σ), the temperature, and other external conditions. Four sequences out of nine appear to be fast folders displaying a two-state kinetic approach to the native conformation with $\Phi \gtrsim 0.9$. These sequences have σ values less than about 0.4.

The other five sequences have Φ values less than 0.9 and hence can be classified as moderate or slow folders. The values of σ for these sequences exceed 0.4. The discrimination of sequences into slow and fast based on Φ is arbitrary. An example of the kinetic behavior of a slow folder (sequence A) probed using $P_u(t)$ is shown in Figure 4b. The generic behavior of $P_u(t)$ as a sum of several exponentials has been argued to be a consequence of KPM [16,21]. Typically, for slow-folding sequences, τ_{FAST} varies from $200\tau_L$ to $600\tau_L$, whereas the largest τ_k (as defined by eq. 26) lies in the interval from $2500\tau_L$ to $2.7 \times 10^6\tau_L$. Slow folding trajectories reach the native state via a TSMM [16,21,24], according to which random collapse of a protein (first stage) is followed by a slow search of the native state among compact conformations (second stage) that eventually leads the polypeptide chain to one of several misfolded structures. These misfolded structures have many characteristics of the native state. Generically, the rate-determining step in the TSMM involves the transition (crossing a free energy barrier) from the misfolded structure to the native state (third stage) [24].

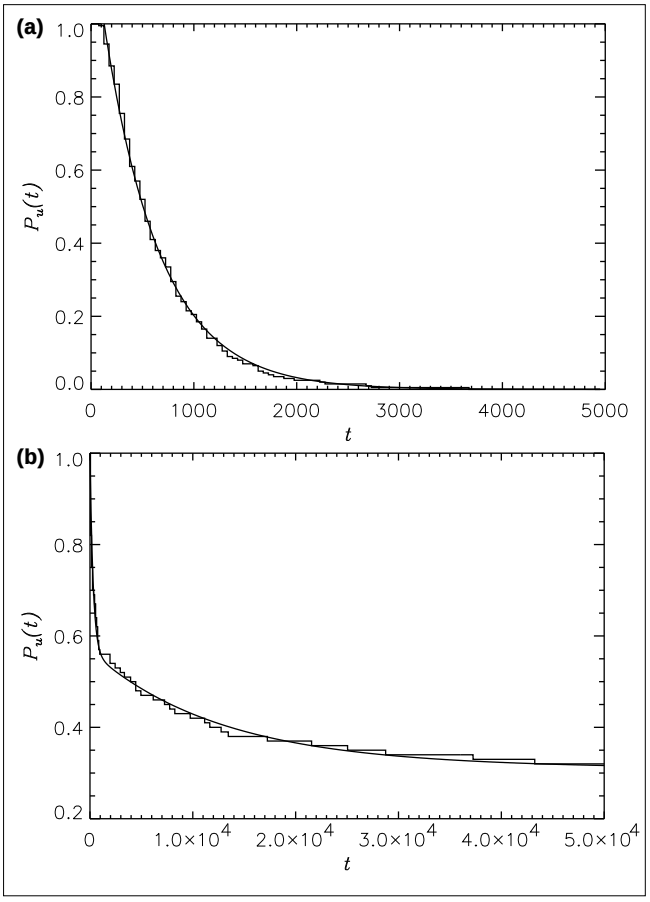
In order to obtain insights into the microscopic origins of the slow and fast phases, we have analyzed the dynamic behavior of various trajectories. We have found that for sequences that find the native conformation in essentially a kinetically two-state manner, all the trajectories reach the native conformation without forming any discernible intermediates. Furthermore, for these cases, once a certain number of contacts is established, the native state is reached very rapidly, which is reminiscent of a nucleation process [6,16,17]. The timescale for such nucleation-dominated processes is relatively short and it has been suggested that in these cases the collapse process and the acquisition of the native conformation occur almost simultaneously [21]. It is for this reason that we refer to this process as native conformation nucleation collapse (NCNC; it has been referred to as a nucleation-condensation mechanism by Fersht [10]). These points are illustrated by examining the dynamics of the structural overlap function $\chi(t)$ for fast folders. A typical plot for $\chi(t)$ for a fast folder (sequence G) is shown in Figure 5 — we plot (in Fig. 5 and later figures) $\chi(t)$ for a trajectory labeled k averaged over a few integration steps h, i.e.:

$$\overline{\chi^k}(t) = \frac{1}{\overline{\tau}} \int_{t-\overline{\tau}/2}^{t+\overline{\tau}/2} \chi^k(s)ds \qquad (29)$$

The value of $\overline{\tau} = 5\tau_L$, which is much less than any relevant folding timescale. Figure 5a shows that within $380\tau_L$ (the first passage time) the chain reaches the native conformation. After the chain reaches the native state there are fluctuations around the equilibrium value of $\langle\chi\rangle$ (= 0.26). Another example of a folding trajectory for this sequence is presented in Figure 5b and is further analyzed below.

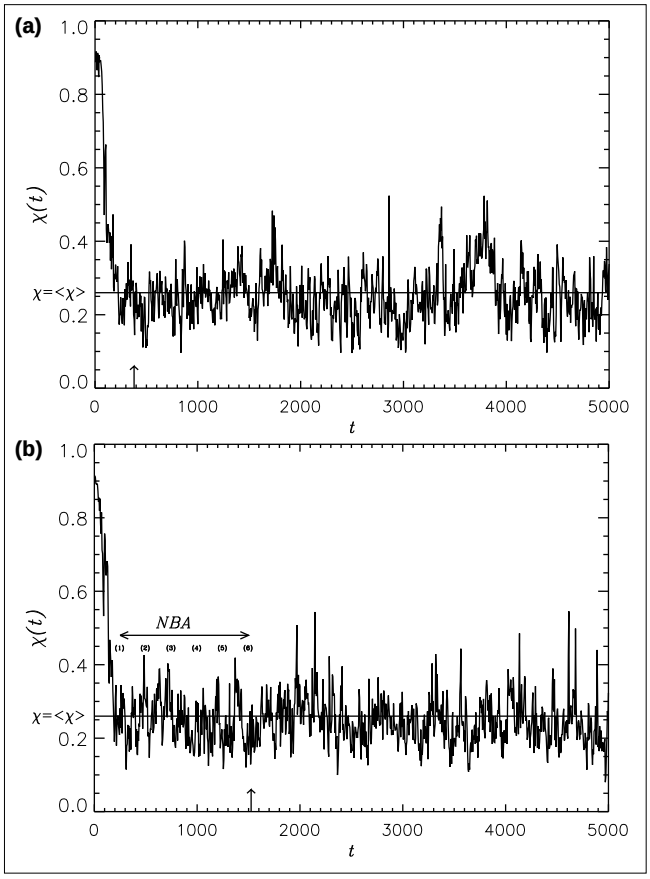
The dynamical behavior shown in Figure 5 for fast trajectories should be contrasted with the trajectories for other sequences that reach the native state by indirect off-pathway processes. An example of such a behavior for the moderate folder sequence E (Φ = 0.72) is shown in Figure 6. The behavior presented in Figure 6a shows that after an initial rapid collapse (on the timescale of about $100–200\tau_L$), the chain explores an intermediate state (where $\chi(t)$ is roughly constant for a large fraction τ_I/τ_{1i} of time, τ_I being the lifetime of the intermediate state) before reaching the native conformation at $\tau_{1i} = 3026\tau_L$. Figure 6b shows another off-pathway trajectory for this sequence, in which the native conformation is reached at $1970\tau_L$. Although these slow trajectories are qualitatively similar, they clearly demonstrate that the chain samples different misfolded conformations depending on the initial conditions before it finally finds the native state. This fact further supports the multipathway character of the indirect folding process. After the native conformation is reached, the overlap function fluctuates around the equilibrium value $\langle\chi\rangle$ = 0.26 or makes sudden jumps to the higher values of $\chi \approx 0.4$ and fluctuates around these values for a finite time. Such dynamics clearly reflects frequent visits to low-lying structures (see below). The behavior shown in Figure 6 is very

Figure 4



The fraction of unfolded molecules $P_u(t)$ as a function of time for (a) sequence G and (b) sequence A. Time is measured in units of τ_L (cf. eq. A1). The solid line in (a) is a single exponential fit to the data. This implies that for this sequence, folding is kinetically a two-state process ($\Phi = 1.0$). The solid line in (b) is a three exponential fit to the data. The multiexponential process is indicative of the kinetic partitioning mechanism with $\Phi = 0.43$ (see eq. 26).

Figure 5



Dynamics of a typical fast-folding trajectory as measured by $\chi(t)$ (sequence G) at ζ_L. (a) In a very short time $380\tau_L$, the native conformation is reached. After the native conformation is reached, $\chi(t)$ fluctuates around the equilibrium value $\langle\chi\rangle$. (b) Another trajectory for this sequence, for which inherent structures at the times labeled 1–6 were determined. Horizontal arrow indicates the region of the native basin of attraction (NBA). It is seen that the chain approaches the NBA but spends a finite amount of time there before reaching the native state at $1525\tau_L$. Horizontal line indicates $\langle\chi\rangle = 0.26$ at T_s. Vertical arrows indicates the first passage time.

typical of the trajectories that reach the native conformation via indirect mechanisms which are conveniently quantified in terms of the TSMM. Figure 7 presents a typical indirect trajectory for fast sequence I, which has the partition factor Φ slightly less than unity. This trajectory reaches the native conformation at $2384\tau_L$.

It is also instructive to compare the dynamical behavior of the nucleation trajectories of fast and slow folding sequences. An example of a trajectory that reaches the native conformation via nucleation collapse mechanism for sequence E is shown in Figure 8. It is important to note that the qualitative behavior of $\chi(t)$ presented in Figure 8 is very similar to that shown in Figure 5. This further confirms that the underlying mechanism that leads the chain directly to the native conformation for sequences with large σ is similar to the nucleation process. The only difference is that the partition factor Φ is less for sequences with large σ than for ones with small σ. Figure 8 also indicates that after reaching the native state, the chain makes frequent visits to neighboring misfolded conformations and, in some instances, gets trapped in these for relatively long times.

The kinetic behavior described above suggests that the value of σ can be used to classify sequences according to their ability to access the native state. It appears that not only does σ correlate well with the intrinsic kinetic accessibility of the native conformation, it also statistically determines the kinetic partition factor Φ. In Figure 9a, we show the dependence of Φ on σ for the nine sequences. The trend that emerges from this plot is that the sequences with larger values of σ (and consequently with larger

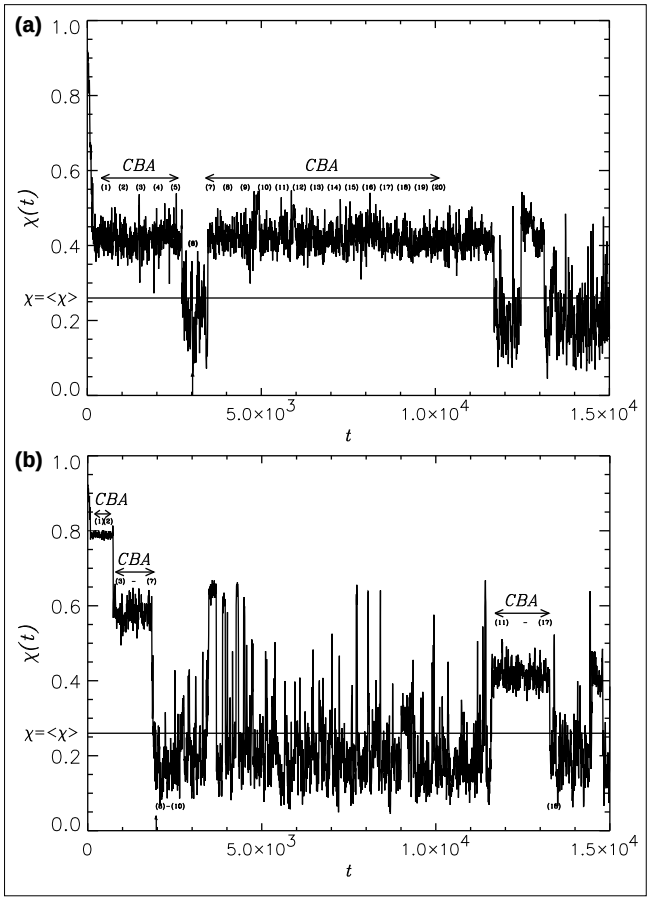
τ_{MFPT}) have smaller values of Φ. For example, for the slow-folding sequence A with the largest value of $\sigma = 0.65$, the fraction of fast trajectories is $\Phi = 0.43$. In contrast, the fastest folding sequence I ($\sigma = 0.14$), for which biexponential fit of $P_u(t)$ is needed, has the value of $\Phi \gtrsim 0.9$.

Probes of kinetic and equilibrium intermediates using inherent structures: roles of native and competing basins of attraction

The issue of the nature and relevance of intermediates in protein folding is of abiding interest. Our studies here and elsewhere [21,43] have demonstrated that the scenarios for folding can be conveniently classified in terms of σ provided the foldable sequences are compared in a similar manner. In order to probe the role of intermediates in the approach to native state, we have analyzed three sequences (E, G, and I) using the kinetic order parameter profiles. Sequences G and I are classified as fast folders (the partition factor Φ exceeds 0.9) whereas sequence E is a moderate folder with the associated σ (Φ = 0.72) lying in the boundary between fast and slow folding sequences.

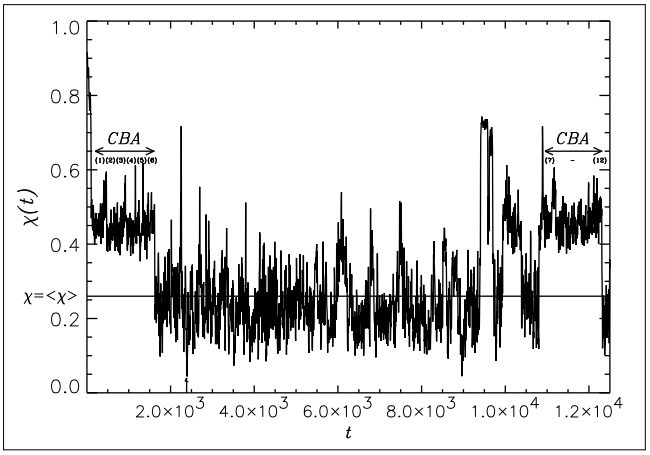
We analyze the role of kinetic and equilibrium intermediates (defined below) using the following methodology. Each trajectory is divided into a kinetic part and an equilibrium part. The kinetic part of a trajectory labeled i includes the portion from the beginning till the native state is reached for the first time, i.e. the first passage time, τ_{1i}. The equilibrium part is taken to be the remaining portion of the trajectory from τ_{1i} till τ_{max}. For convenience, we take τ_{max} to be the same for all trajectories. In order to characterize the nature of intermediates we use the overlap function, χ, which as described earlier gives the degree of similarity to the native conformation. It is possible that the same value of χ may correspond to different conformations and in some instances to conformations that are even structurally unrelated to each other. However, by studying the distribution of overlap function over a range of χ for several independent initial conditions and by directly comparing the resulting conformations and calculating χ between them, we can ascertain the states that are visited with overwhelming probability before and after reaching the native conformation. In order to probe the nature of kinetic and equilibrium intermediates that the chain samples en route to the native conformation, we have determined the 'inherent' structures [53] (see below). The inherent structures are obtained from the timecourse of $\chi(t)$, examples of which are shown in Figures 5–8. The basins of attractions are obtained before the chain reaches the native conformation for the first time (i.e. the 'kinetic' basins) and are determined as follows. As the polypeptide chain approaches the native conformation (but has not yet reached it, i.e. $t < \tau_{1i}$), we record several (usually about 10) instantaneous conformations which serve as initial conditions for steepest descent simulations. In this method, the temperature is set to zero and the velocities of all residues are rescaled to zero after each integration step. This results in a 'downhill' motion of a sequence on the energy surface. The final conformations of the steepest descent quench simulations (provided they are sufficiently long) are the conformations of local energy minima (inherent structures)

Figure 6



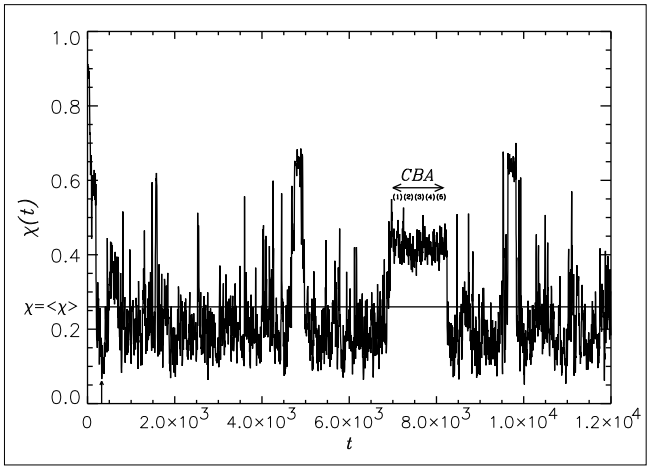
Examples of two trajectories that reach the native conformation by an indirect off-pathway process. These trajectories are for sequence E at ζ_L. The kinetics exhibited by the off-pathway process suggests that the native state is reached by a three-stage multipathway mechanism. **(a)** After initial rapid collapse on the timescale of $100–200\tau_L$, the chain gets trapped in misfolded compact structure (indicated by a nearly constant value of χ for long times). In this case, the native state is eventually reached at $\approx 3026\tau_L$. **(b)** A different slow trajectory showing that the chain samples at least two distinct misfolded structures before the first passage time is attained at $\approx 1970\tau_L$. In both parts, numbers indicate the points at which inherent structures were determined. The timecourse of $\chi(t)$ reveals that the chain samples a number of kinetic and equilibrium intermediates. It was found that inherent structures 1–5 and 7–20 in (a), 11–17 in (b), and 1–5 in Figure 8 are almost identical and are accessible before and after first passage time. For this, they are classified as native-like equilibrium intermediates. However, the inherent structures 1 and 2 and 3–7 in (b) are examples of kinetic intermediates. Horizontal lines in these plots indicate the equilibrium value of $\langle\chi\rangle = 0.26$ at T_s. Horizontal arrows indicate the regions of competing basins of attraction (CBAs). Vertical arrows indicate the first passage time.

Figure 7



Dynamics of one of the few off-pathway trajectories for sequence I ($\Phi = 0.95$). Inherent structures are determined at the points marked by the numbers. Analysis shows that structures 1–6 and 7–12 are identical, which allows us to refer to them as equilibrium native-like intermediates. Note that the vast majority (≈ 0.95) of trajectories fold via a native conformation nucleation collapse (NCNC) mechanism. The native state is reached at $2384\tau_L$. Horizontal arrows indicate the regions of CBAs. Vertical arrow indicates the first passage time.

Figure 8



The dynamics of $\chi(t)$ for one trajectory for sequence E ($\Phi = 0.72$) that reaches the native conformation rapidly. In this example, the native conformation is attained at $\approx 325\tau_L$. Comparison with Figure 5 (for sequence G with $\Phi > 0.9$) shows that the dynamics is very similar. This implies that the underlying mechanism (NCNC) of fast-folding trajectories of sequences with large σ (or equivalently small Φ) is similar to that by which the molecules reach the native state in a kinetically two-state manner. Horizontal arrow indicates the regions of CBAs. Vertical arrow indicates the first passage time.

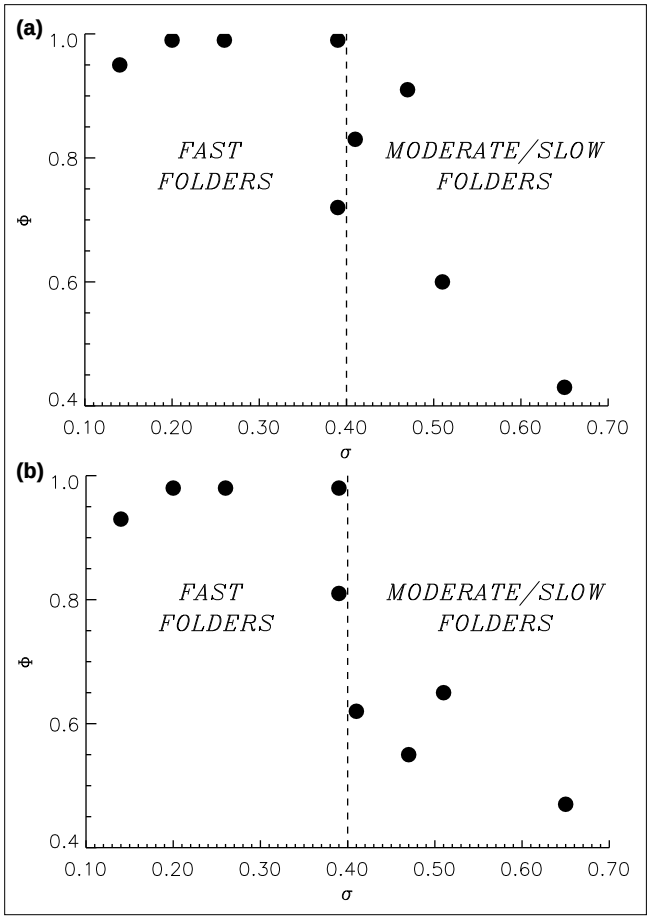
which the sequence explores in the folding process. These conformations obtained at different times and with distinct initial conditions allow us to map the distribution of folding pathways. The same technique for getting inherent structures was also used after the first passage time $\tau_{1i} < t < \tau_{max}$. These give us the 'equilibrium' intermediates. This analysis allows us to compare the nature of intermediates in the kinetic pathways.

For sequence G, we determined the inherent structures using the instantaneous conformations labeled 1–6 (all of which occur at $t < \tau_{1i}$) in Figure 5b. The inherent structures for this particular trajectory (and for others) almost always coincide with the native state. This clearly shows that for sequence G, for which the native state is reached by the nucleation collapse mechanism, the various inherent structures directly map into the NBA. The rapid approach to the NBA is the reason for the two-state kinetics displayed. It also follows that the NBA is relatively smooth, i.e. the energy fluctuations characterizing the roughness are comparable to $k_B T_s$. The roughness associated with the NBA implies that the polypeptide chain spends a finite amount of time in close proximity ($\chi(t) \approx \langle\chi\rangle$) to the native conformation prior to reaching it. It is worth emphasizing that this sequence ($\sigma = 0.20$) fluctuates around only the native state even for $t > \tau_{1i}$ for all the trajectories examined.

Of the nine sequences we have examined, I is the fastest folder, i.e. has the smallest folding time. Nevertheless, the partition factor Φ is slightly (but measurably) less than unity. The amplitude of the slow component is very small (for this sequence the biexponential fit to $P_u(t)$ suffices). These observations suggest that the underlying topography explored could be somewhat different from that of sequence G which is also a fast folder. Most of the trajectories reach the native state for sequence I rapidly without forming any intermediates and resemble the behavior (shown in Fig. 5) for sequence G. However, there are 'off-pathway' trajectories for this sequence (an example of which is shown in Fig. 7). The inherent structures at the kinetic part for this particular trajectory $t < \tau_{1i}$ were determined using the conformations labeled 1–6 in Figure 7. In addition, the inherent structures were also calculated using the conformations 7–12 that the chain samples after the first passage time for this trajectory. We found that these inherent structures are all identical and differ very slightly (as measured by the overlap function). Consequently, we characterize them as native-like intermediates. This sequence, although a fast folder, has at least one competing basin of attraction (CBA) in which the structure is quite similar to the native state. As a small fraction of molecules reach the CBA prior to reaching the NBA, the Φ value is smaller than unity. The comparison between sequences G and I, both of which fold very rapidly, shows that there can be significant differences in the underlying energy surface (further illustrated below).

According to our classification, sequence E is at least a moderate folder and exhibits the full range of the KPM (Φ

Figure 9



Correlation between the fraction of fast-folding trajectories Φ and the parameter $\sigma = (T_\theta - T_F)/T_\theta$. Most sequences with small σ have $\Phi \approx 1.0$. Vertical dashed lines show the classification of sequences with respect to Φ. (a) Low friction value. (b) Moderate friction. Sequences with $\Phi \gtrsim 0.9$ are classified as fast folders. The classification of sequences as slow folders is somewhat arbitrary. The classification does not seem to depend on the value of ζ.

= 0.72). A significant component of initial trajectories reach the native state via TSMM (examples of these off-pathway trajectories are shown in Fig. 6). We have obtained the inherent structures using conformations labeled 1–6 in Figure 6a (that occur before τ_{1i}) and using the conformations labeled 7–20 (that are sampled for times greater than τ_{1i}). It is found that these structures are nearly the same (excluding structure 6) indicating that, in this instance, the polypeptide samples native-like intermediates en route to the native conformation. In this sense, the behavior for this trajectory is no different from that observed for off-pathway trajectories for sequence I (Fig. 7).

The result of a similar analysis using another trajectory (shown in Fig. 6b) is dramatically different. The inherent structures obtained using the conformations labeled 1 and 2 and 3–7 are completely different from each other. Furthermore, the equilibrium intermediates identified with the inherent structures obtained using the instantaneous conformations 11–17 do not resemble those calculated during the kinetic portion 1–7. We do find that the equilibrium intermediates for this trajectory 11–17 are virtually identical to those calculated using the conformations (CBAs) sampled by other trajectories (displayed in Figs 6a and 8). Examination of other off-pathway trajectories reveals the presence of an exceptionally stable intermediate with $\chi \approx 0.8$. In fact, this intermediate survives for $90\,000\tau_L$, while a typical first passage time is only about $1000\tau_L$. Such intermediates described above are never visited again after folding is completed, hence they are kinetic intermediates.

These observations imply that for moderate and slow folders there are several CBAs. Some of these serve as equilibrium intermediates, i.e. have native-like characteristics and the chain revisits them even after reaching the native state. Others, which occur relatively early in the folding process, perhaps during the initial collapse process itself, are kinetic intermediates that are not visited after the native state is reached, at least during the timecourse of our simulations. Thus, for moderate and slow folders, one has a distribution of CBAs. The presence of CBAs provides the entropic barriers to folding [50] resulting in a slow approach to the native state. In contrast, for fast folders, the only intermediates that are encountered, if any at all, are all native-like. Thus, for fast-folding sequences only the NBA dominates. In such cases, the energy landscape can be thought of as being funnel-like [2,28].

Free energy profiles

The analysis in the preceding subsection indicates that the free energy profile can be quite complex. The shapes of these profiles depend crucially on the sequence and external conditions (in our simulations that is specified only by the temperature). We have attempted a caricature of the free energy surface by computing the histogram of states expressed in terms of the potential energy E_p and χ. The histogram of states, which measures the probability of occurrence of the state with a given E_p and χ, is defined as:

$$g(E_p,\chi) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{\tau_{max}-\tau_{1i}}\int_{\tau_{1i}}^{\tau_{max}}\delta(E_p - E_{p,i}(t))\,\delta(\chi-\chi_i(t))ds \qquad (30)$$

where $E_{p,i}(t)$ and $\chi_i(t)$ are the values of potential energy and overlap function for the trajectory i at time t averaged over a small interval of $5\tau_L$. We have calculated $g(E_p,\chi)$ for three sequences at the sequence-dependent simulation temperature T_s. The values of $M = 100$, and a grid size of 0.1 used for E_p and χ is increased in intervals of 0.01. If $\tau_{max} >> \tau_{1i}$ then eq. 30 gives the equilibrium distribution

function. A free energy profile may be illustrated using the potential of mean force defined as:

$$W(E_p,\chi) = -k_B T_s \ln\left[g(E_p,\chi)\right] \quad (31)$$

In Figures 10–12, we plot $g(E_p,\chi)$ for the three sequences G, I and E and show the contour plot of the histogram of states. For sequence G (Fig. 10), it is clear that the NBA is the only dominant maximum and consequently the kinetics on this surface is expected to be two-state-like. The plots for sequence G also show that after the NBA is located, the chain fluctuates only in the NBA. The free energy profile for sequence I (as suggested by Fig. 11) has in addition to the NBA at least one CBA. The presence of the CBA makes Φ smaller than that for sequence G (for which Φ = 1.0). Proteins with a larger σ would have several CBAs. This is clearly indicated in Figure 12 for sequence E, showing two discernible CBAs, which makes this model protein only a moderate folder. The profile of the potential of mean force for this sequence, computed using eq. 31, is shown in Figure 13. This figure shows that in general one has a complex structure for the free energy profile. It is also clear that this multivalley structure naturally leads to the KPM (discussed above). These figures also show that in special cases (small values of σ), the folding kinetics can be described in terms of only the NBA or folding funnel [2,28].

Dependence of τ_F on σ

It is clear from the results discussed above that the parameter σ (for a given external condition, which in our case is the simulation temperature) may be used to predict approximate kinetic behavior of various sequences. The folding time τ_F, which is taken to be the mean first passage time τ_{MFPT}, is plotted as a function of σ in Figure 14a. This graph shows a remarkable correlation between τ_F and σ. The sequences with small $\sigma \lesssim 0.4$ fold to the native conformation very rapidly, so that τ_F is less than about $600\tau_L$. However, τ_F for the sequence with largest σ = 0.65 is as large as $875\,258\tau_L$. Thus, variation of the parameter σ from 0.14 to 0.65 results in three orders of magnitude increase in the folding time (from $461\tau_L$ to $875\,258\tau_L$). It must be noted that the correlation between τ_F and σ should be considered as statistical. One can easily notice a few pairs of closely located data points in Figure 14a for which a larger value of σ does not correspond to larger τ_F. Nevertheless, the general conclusion remains apparent: the parameter σ allows us to predict the trend in the folding rate of the sequences by knowing only its thermodynamic properties, such as T_θ and T_F. It should also be pointed out that because of the difficulty in computing the low energy spectra of the off-lattice models ([27]; M Fukugita, personal communication), correlations between folding times and other quantities (such as the energy gap or the relative value of the native energy compared to that of nonnative conformations) were not tested. In addition, there appears to be no unambiguous way to determine the kinetic glass transition temperature, $T_{g,kin}$. Therefore, we have not tested the proposal that foldable sequences have large values $T_F/T_{g,kin}$ [54].

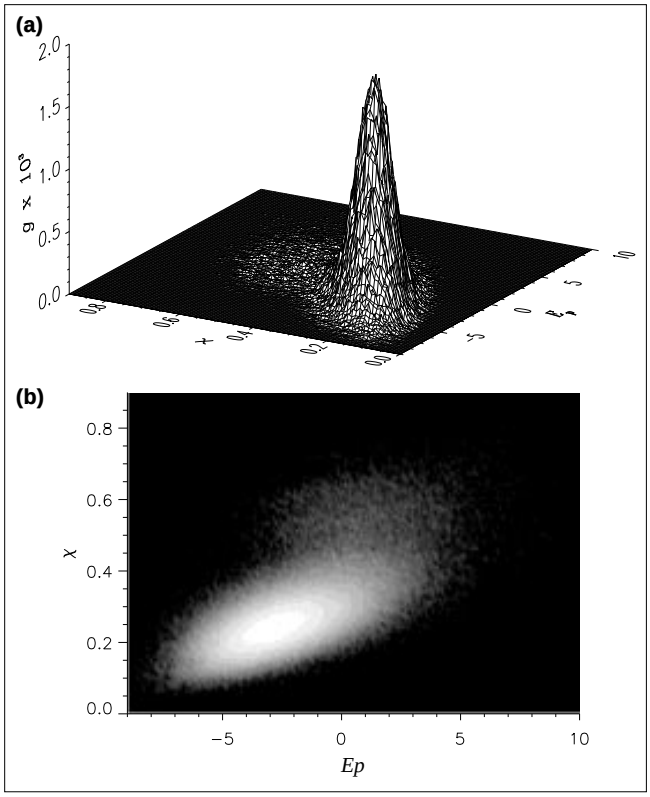
In order to study the dependence of the folding time on the parameter σ, we used the function $P_u(t)$ and defined folding time as the mean first passage time τ_{MFPT} (see eq. 24). As mentioned above, the alternative is to analyze the overlap function $\langle\chi(t)\rangle$ and take the largest exponent τ_2 in the exponential fit to $\langle\chi(t)\rangle$ as an estimate for the folding time. Due to computational limitations we did this for only five sequences and found the trend to be similar to that illustrated in Figure 14a, i.e. the folding time τ_2 correlates remarkably well with the parameter σ.

Kinetics and folding times at moderate friction

The results presented above were obtained with the value of friction coefficient fixed at ζ_L = 0.05. In order to study the dependence of the folding kinetics on ζ, we have performed the same study of nine sequences at a larger value of the friction coefficient $\zeta_M = 5 = 100\zeta_L$. The plot showing the folding time $\tau_F = \tau_{MFPT}$ as a function of the parameter σ at ζ_M is displayed in Figure 14b. In accord with the results obtained at the lower value of ζ_L, this also unambiguously demonstrates a good correlation between σ and τ_F, so that the sequences with small values of σ fold much faster than the sequences having large σ. Specifically, sequence I, which has the smallest value of σ = 0.14, reaches the native conformation very rapidly within $\tau_F = 1554\tau_L$, while sequence A with σ = 0.65 folds very slowly within $\tau_F = 2.4 \times 10^6\tau_L$. As one may expect, the overall folding times in the moderate friction limit are considerably larger than in the low friction limit. The folding times vary almost linearly with ζ. For most sequences, the ratio $\tau_F(\zeta_M)/\tau_F(\zeta_L)$ is 3–4. The largest value of this ratio is found for the slow-folding sequence D and is equal to 5.

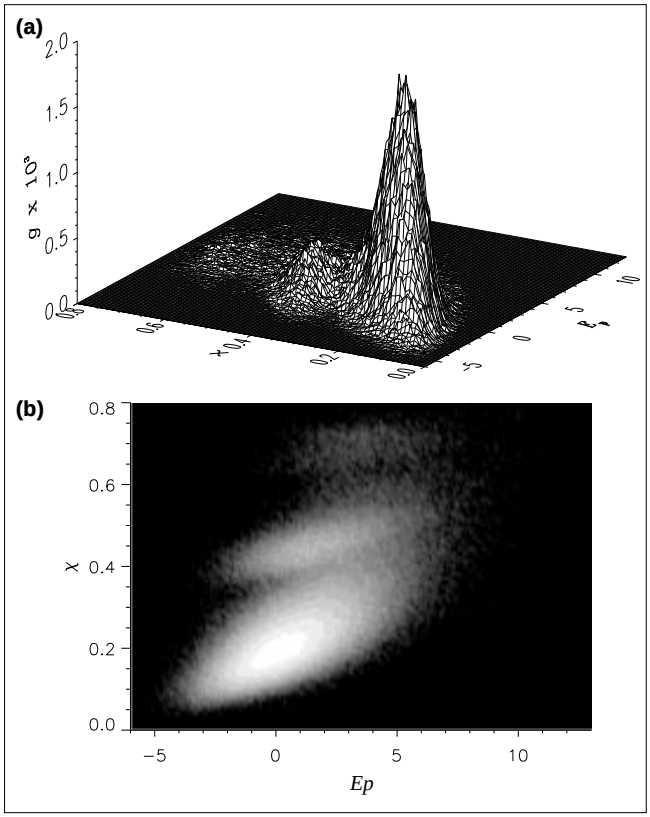
In order to compare the folding kinetics at ζ_M with those obtained at ζ_L we analyzed several folding trajectories. Figure 15 presents typical folding trajectory (in terms of the overlap function $\chi(t)$) for sequence G, which displays two-state kinetics and is classified as a fast folder. After a few tertiary native contacts are established, the chain rapidly reaches the native state. In Figure 16, we plot $\chi(t)$ for typical slow (Fig. 16a) and fast (Fig. 16b) trajectories for sequence E, which, in contrast to sequence G, exhibits KPM and is classified as a moderate folder. It is seen that the fast trajectory for sequence E is very similar to a typical trajectory for sequence G. The reason for this is that the underlying mechanism for the fast process, NCNC, is exactly identical. It is also very important to note that similar plots for these two sequences (Figs 5–6,8) obtained at ζ_L are virtually the same as those shown in Figures 15 and 16. This allows us to suggest that principal mechanisms of protein folding, such KPM and

Figure 10



(a) Histogram of states g as a function of two variables, E_p and χ, for sequence G. (b) The contour plot of the histogram of states g for this sequence. Lighter areas correspond to peaks of g. The single peak of the histogram of states suggests that at equilibrium the chain is completely confined to a native basin of attraction.

Figure 11



(a) Histogram of states g and (b) contour plot of g for sequence I. These plots reveal two peaks of the histogram of states that manifest the presence of a competing basin of attraction that makes the partition factor Φ less than unity.

nucleation collapse, appear to be independent of the viscosity of surrounding medium. The timescales, however, depend critically on viscosity [21].

The classification of sequences into slow and fast folders based on the parameter σ can also be carried out with the larger value of ζ_M. Fast-folding sequences (four out of nine) are characterized by values of $\sigma \gtrsim 0.4$. The mean first passage time for fast folders τ_{MFPT} is below $3000\tau_L$. The function $P_u(t)$ for fast folders is adequately fit (apart from one sequence) with a single exponential just as in the low friction limit. Thus, folding of these sequences proceeds via a nucleation collapse mechanism. The sequences with $\sigma \gtrsim 0.4$ can be classified as slow or moderate folders. These sequences have significantly larger mean first passage times τ_{MFPT} ranging from $3285\tau_L$ to $2.4 \times 10^6\tau_L$. Most importantly, the fraction of unfolded molecules $P_u(t)$ is clearly two or three exponential (see eq. 26) which is an apparent manifestation of KPM. As for the low friction limit, the fraction of fast-folding trajectories Φ increases as the parameter σ decreases (Fig. 9b). Specifically, for sequence A (σ = 0.65), Φ = 0.47, while for the fastest folding sequence I (σ = 0.14), the fraction of fast trajectories becomes as large as 0.93.

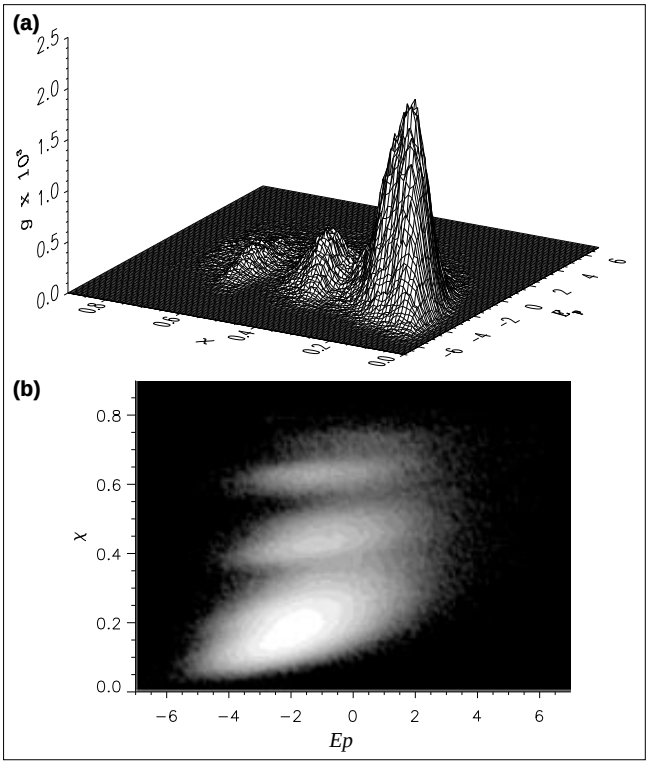
Quantitative dependence of τ_F on σ

It is interesting to comment on the quantitative dependence of τ_F on σ. Theoretical arguments suggest that, at least at small values of σ, τ_F should scale algebraically with σ, i.e. $\tau_F \sim \sigma^\theta$ with $\theta = 3$ [21]. The present simulations as well as previous studies using lattice models [23,51] suggest that the data can also be fit with an exponential, i.e.:

$$\tau_F \simeq \tau_0 F(N) \exp\left(\frac{\sigma}{\sigma_0}\right) \qquad (32)$$

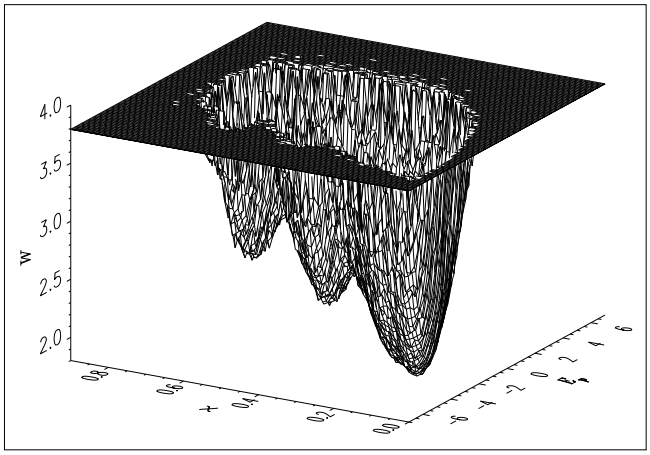
where $F(N)$ is a function that depends on N. It has been argued [21] that $F(N) \sim N^\omega$ with $3.8 \lesssim \omega \lesssim 4.2$ for $\sigma \approx 0$ and $F(N) \sim \exp(\sqrt{N})$ for larger σ. The data in Figure 14 can be fit with eq. 32 with $\sigma_0 \approx 0.06$ at ζ_L and ζ_M. The fit of τ_F to an algebraic power ($\tau_F \sim \sigma^\theta$) gives $\theta \approx 3.9$ at ζ_L and ζ_M. Further work will be needed to fully quantify the precise dependence of τ_F on σ. It appears that both eq. 32 and the algebraic behavior [21] account adequately for the

Figure 12



(a) Histogram of states *g* and (b) contour plot of *g* for sequence E. One can clearly see at least three maximums of *g*. These plots illustrate the existence of several competing basins of attraction (intermediates) that gives rise to complex folding kinetics which feature a combination of three-stage multipathway and nucleation collapse mechanisms.

Figure 13



The profile of the potential of the mean force *W* in terms of two variables, E_p and χ, for sequence E. This further illustrates that the free energy landscape of this sequence features multiple funnels (basins of attractions). The plane at *W* = 3.8 is given for eye reference.

data given here and elsewhere for lattice models. The fit given in eq. 32 appears to be a bit more accurate.

Implications for experiments

The results presented here, together with the timescale estimates given in the Appendix, have a number of implications for experiments. Here, we restrict ourselves to providing some comparisons to the folding of chymotrypsin inhibitor 2 (CI2) which was probably the first protein for which a kinetic two-state transition was established [26,55]. These experiments established that the kinetics for the fast phase, which corresponds to the molecules with proline residues in a *trans* conformation, follows a two-state behavior. Furthermore, the thermodynamics also displays a two-state cooperative transition with the native conformation being stable by about 7 kcal mol^{-1} at T = 25°C, pH = 6.3 and at zero denaturant concentration. Although not explicitly addressed here, we have argued elsewhere [21] that the marginal stability (relative to other structurally unrelated conformations) of the native state of proteins satisfies:

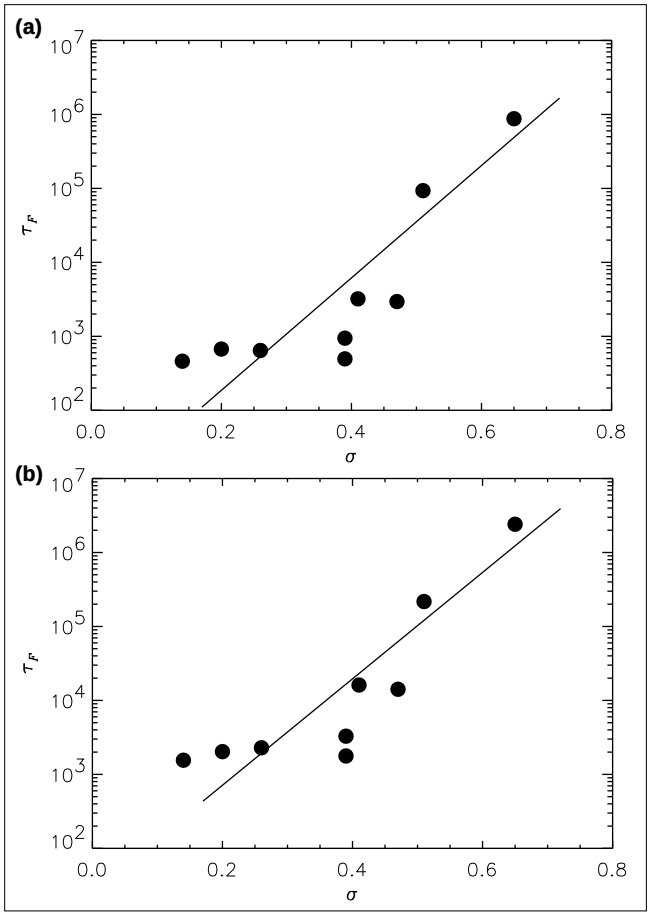
$$\Delta G \gtrsim k_B T\sqrt{N} \qquad (33)$$

where the unknown prefactor is assumed to be of the order of unity. The CI2 examined by Jackson and Fersht [26,55] has 83 residues and consequently eq. 33 gives $\Delta G \approx$ 5.5 kcal mol^{-1} at T = 25°C. This is in fair agreement with the experimental determination. It appears that eq. 33 is consistent with the marginal stability of proteins of varying size. The bound given above seems to be a good estimate of the stability of biomolecules [8]. We expect the scaling relation of the type given in eq. 33 to be accurate to only within a factor of two. Given the inherent experimental uncertainty in determining ΔG, the agreement with the theoretical prediction within ~20% is remarkable.

The kinetics of folding of CI2 can be rationalized using the ideas developed here. The timescale for NCNC according to eq. A5 is $\tau_{NCNC} \approx 0.2$ ms using the parameters specified in the Appendix and with $\sigma \approx 0.4$ (we have taken $T_\theta \approx 60$°C and $T_F \approx 37$°C). If we assume that the folding time changes exponentially with σ (cf. eq. 32), then the estimate for the nucleation collapse time changes to about 25 ms, where we have used $\sigma_0 \approx 0.1$. These estimates give an interval (a relatively broad one) 0.2 ms $\lesssim \tau_{NCNC} \lesssim$ 25 ms. Despite the uncertainties in the theoretical estimates (unknown prefactors, errors in the estimates of γ, a_0, etc.), the estimated values of τ_{NCNC} are within measured experimental values. The early experiments and more recent ones on CI2 and a mutant of CI2 indicate that the folding time for τ_{NCNC} is in the range 1.5–18 ms [9,15,26,55].

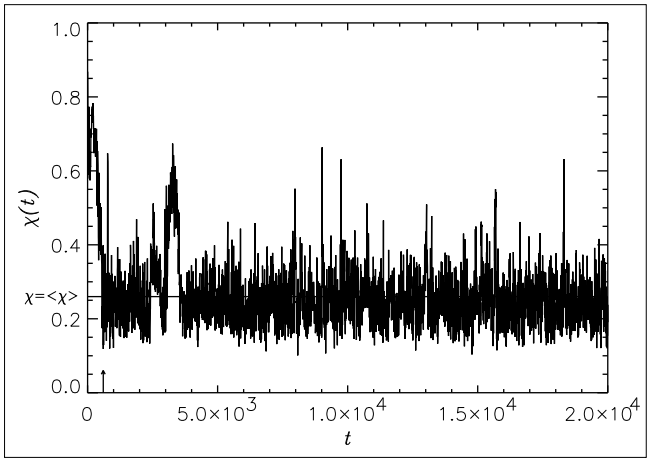
The fastest folding time of 1.5 ms is found for a mutant of CI2 (DE Otzen, AR Fersht, personal communication). Our theoretical estimates show that even if the external conditions are constant and the length of the polypeptide chain is fixed, τ_{NCNC} can still be altered if σ (see eq. A5) is

Figure 14



Dependence of the folding time τ_F on the parameter $\sigma = (T_\theta - T_F)/T_\theta$. It is seen that τ_F correlates remarkably well with σ, so that sequences with small values of σ reach the native state very rapidly, whereas those characterized by large σ fold slowly. Solid lines indicate the exponential fit of the data. The actual fit to the data is discussed in the text. (a) Low friction limit. (b) Moderate friction limit.

Figure 15



An example of a fast-folding trajectory that reaches the native conformation by a NCNC process. This is for sequence G at ζ_M. The dynamics of the fast-folding trajectory is qualitatively similar to that obtained at ζ_L (see Fig. 5). In both cases, the native conformation is reached rapidly following the formation of a critical number of contacts (nucleus) and collapse. The first passage time for this trajectory is $603\tau_L$. Horizontal line gives the equilibrium value of $\langle\chi\rangle = 0.26$ at T_s. Vertical arrow indicates the first passage time.

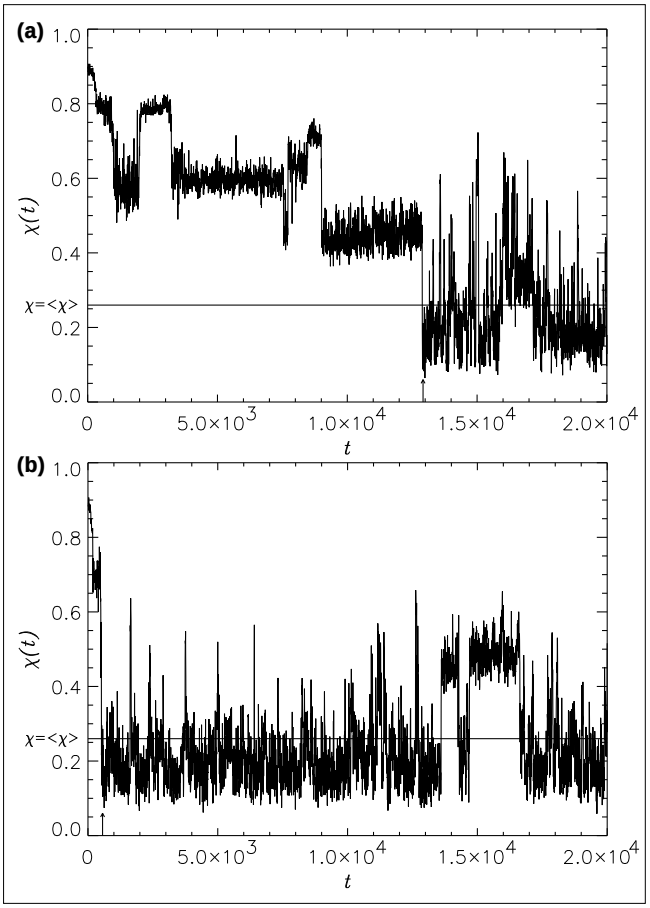
altered. Since σ is very sensitive to sequence, we suggest that the mutant of CI2 has a different value of σ than the wild type. This can readily explain the decrease in folding time for the mutant under otherwise similar external conditions. Further work is needed to quantify these ideas.

Conclusions

The folding of proteins is a complex kinetic process involving scenarios that are not ordinarily encountered in simple chemical reactions. This complexity arises due to the presence of several energy scales and the polymeric nature of polypeptide chains. As a result, this complexity leads to a bewildering array of timescales that are only now beginning to be understood quantitatively in certain minimal models of proteins [21,22]. Despite this remarkable complexity it has been known from the pioneering studies of Anfinsen that the specification of the primary sequence determines the three-dimensional structure of proteins, i.e. native state topology is encoded in the primary sequence. The study presented here and our earlier work on lattice models [23,43] have shown clearly how the kinetic accessibility is also encoded in the primary sequence itself. Our results suggest that a wide array of mechanisms that are encountered in the folding process are, remarkably enough, determined by a simple parameter expressible in terms of the properties that are intrinsic to the sequence but affected by external conditions. It appears that the two characteristic equilibrium temperatures T_θ and T_F determine the rate at which a given sequence reaches the native conformation. T_θ and T_F not only depend on the sequence but also can be dramatically changed by varying the external conditions such as pH, etc. Thus, the mechanism for reaching the native conformation for a single domain protein can change dramatically depending on the external conditions. This implies that a protein that exhibits two-state kinetics under given external conditions does not necessarily follow the same kinetics if the ambient conditions (e.g. pH) are altered.

Our results show that generically the polypeptide chain reaches the native conformation by a kinetic partitioning mechanism. For a number of sequences studied here we have established that for given external conditions (for the computational studies it is the temperature only), a fraction of molecules Φ reaches the native conformation directly via nucleation collapse mechanism, while the remainder follows a complex three-stage multipathway kinetics. For both values of friction coefficient studied here this general scenario holds.

Figure 16



(a) An example of a slow-folding trajectory as recorded by $\chi(t)$ (sequence E) at ζ_M. After initial rapid collapse on the timescale $\lesssim 1000\tau_L$, the chain samples various compact conformations and finally reaches the native state at $12\,919\tau_L$ as indicated by an arrow. This trajectory shows that at least four distinct kinetic structures are sampled as the chain navigates to the native conformation. (b) Typical fast-folding trajectory for this sequence. This trajectory is similar to those characteristic of fast folders (see Fig. 15). The native conformation is found very rapidly at $563\tau_L$ as indicated by an arrow. The results displayed in Figures 5–8, 15 and 16 show that the qualitative aspect of the kinetic partitioning mechanism is not dependent on the friction coefficient. Horizontal line indicates the equilibrium value of $\langle\chi\rangle = 0.26$ at T_s. Vertical arrow indicates the first passage time.

It is clear from our results that once the external conditions are specified, Φ is essentially determined by the interplay of T_θ and T_F as embodied in eq. 1. The folding time correlates extremely well with the dimensionless parameter $\sigma = (T_\theta - T_F)/T_\theta$ independent of the value of the external friction. The remarkable correlation between σ and several kinetic properties lends credence to the notion that, in small proteins at least, a single collective coordinate description of folding may suffice [56]. It also follows from this study that only when σ is small can folding be described in terms of NBA. For moderate and slow folders it is important to consider the interplay between NBA and CBA in determining folding kinetics. The independence of our general conclusions on the type of models (lattice versus off-lattice) [23,43] and on the details of the dynamics (Langevin or Monte Carlo) seems to indicate that the KPM (along with σ determining the trends in folding times) may describe in a concise fashion the scenarios by which single domain proteins reach the native conformation.

There are quantitative differences between the results obtained for lattice and off-lattice models. For example, using simulations of lattice models it was concluded that fast folders (with $\Phi \approx 1.0$) have values of σ less than about 0.15 [43]. The off-lattice models suggest that fast folders can have σ as large as about 0.4. Since the estimates of T_θ and T_F using the off-lattice simulations appear to be in better accord with experiments, it is tempting to suggest that for semi-quantitative comparison with experiments it is better to use off-lattice simulations.

Appendix

In this appendix, we map the natural time units to real times so that an assessment of the folding times for these minimal models as well as for small-sized proteins can be made. In addition, using a mapping between these models and proteins, estimates for folding times for proteins with a small number ($\lesssim$70) of amino acids are also presented. We expect these estimates to be accurate to within an order of magnitude due to large uncertainties in the estimates of various quantities as well as a lack of theoretical understanding of the conjectures. From the equation of motion (see eq. 11) it is clear that when the inertial term dominates, the natural unit of time is $\tau_L = (ma^2/\epsilon_h)^{1/2}$. Typical values of m_0 and a_0 for amino acid residues are 3×10^{-22} g and 5×10^{-8} cm, respectively. These are the masses and the Van der Waals' radius of the amino acid residues. The hydrophobic interaction energy ϵ_h is of the order of 1 kcal mol^{-1} or (7×10^{-14} erg). If these values are changed by a factor of two or so there will be not a significant change in our conclusions. Assuming that a bead in our model roughly represents one amino acid, we evaluate τ_L as:

$$\tau_L = \left(\frac{m_0 a_0^2}{\epsilon_h}\right)^{1/2} \approx 3\,ps \tag{A1}$$

The value of the low friction coefficient used in our simulations $\zeta_L = 0.05\ m\ \tau_L^{-1} = 5 \times 10^{-12}$ g s^{-1}, while the value of $\zeta_M = 100\zeta_L = 5 \times 10^{-10}$ g s^{-1}. It is interesting to compare these values for ζ to that obtained in water which has at room temperature T = 25°C a viscosity of 0.01*Poise* with 1*Poise* being equal to 1 g s^{-1} cm^{-1}. The friction on a bead of length a_0 may be estimated as:

$$\zeta_{water} \simeq 6\pi\eta_{water} a_0 \approx 9 \times 10^{-9}\ \mathrm{g\ s^{-1}} \tag{A2}$$

From this, we get $\zeta_L/\zeta_{water} \approx 10^{-3}$, while $\zeta_M/\zeta_{water} \approx 0.1$. The low friction would correspond to the energy diffusion regime in the Kramer's description of the unfolding to folding reaction. In the moderate friction there could be a competition between inertial and viscous damping terms leading perhaps to the Kramer's turnover regime familiar in literature on simple reactions.

In the overdamped limit, the inertial term can be ignored and the natural measure for time is:

$$\tau_H \simeq \frac{\zeta a^2}{k_B T_s} \simeq \frac{6\pi\eta a^3}{k_B T_s} = \alpha\tau_L \frac{\epsilon_h}{k_B T_s} \tag{A3}$$

where α is a constant. In our simulations, α = 0.05 for ζ_L and α = 5.0 for ζ_M. The typical value of ϵ_h/k_BT_s is about 2, where once again we have used ϵ_h = 1 kcal mol^{-1} and taken T_s to be room temperature. For water at room temperature $\tau_H \approx 3$ ns with $\alpha \approx 100$. If we assume that the higher value of friction used in this study is in the slightly overdamped limit, then the approximate time unit becomes $\tau_M \approx 0.3$ ns. Since the higher value of friction is more realistic, we can estimate the folding times for small proteins using the computed timescale. For α = 5.0, our simulation results give the folding time ranging from 100 τ_M to 10^6 τ_M. The folding time for the case of higher friction (with α exceeding 5) also ranges from $10^3\tau_M$ (for the smallest σ) to $10^6\tau_M$ (for the largest σ) (DK Klimov, D Thirumalai, unpublished data). A naive estimate using these results would suggest that the folding time for these sequences ranges from 10^{-6} s^{-1} to 10^{-4} s^{-1}.

A better estimate of these times can perhaps be made by recognizing that each bead in the minimal model corresponds to a blob containing g number of amino acids [41]. If the structure within a blob is represented by roughly spherical size a than we can use $a \approx g^{\nu}a_0$, where g^{ν} is the 'swelling factor' mapping the minimal model to real proteins. Then the natural time unit for the motion of such a blob in the overdamped limit becomes:

$$\tau_H^R \simeq \frac{6\pi\eta g^{3\upsilon} a_0^3}{k_B T_s} = g^{3\upsilon}\tau_H \qquad \text{(A4)}$$

The range of ν is 1/3–1, with ν = 1/3 corresponding to globular structure within a blob and ν = 1 corresponding to maximum repulsion among the residues in a blob. This should be viewed as a guess and is not expected to be correct given that g is small. Realistic values of g are expected to be between 2 and 3 ([22]; JD Bryngelson, personal communication) making the 22-mer minimal model to (perhaps) correspond with 44–66 amino acid residue proteins. For g = 2, τ_H^R ranges from 0.4 ns to 1.6 ns, while for g = 3, τ_H^R ranges from 0.4 ns to 5.4 ns. Assuming that g = 3 and ν = 1 (which would give the largest timescales), the folding estimates for small proteins (number of amino acids < 70) range from 8×10^{-6} s^{-1} (for small σ) to 10 ms (for large σ).

This exercise suggests that no matter how the mapping is done, the most relevant timescale for folding kinetics of small proteins under normal folding conditions (around room temperature and low denaturant concentration) is between microseconds to milliseconds. In particular, for those proteins that reach the native conformation predominantly via the nucleation collapse process (characterized by relatively small σ), the timescale for folding is between microseconds to milliseconds for small proteins. One of us has argued [21] that the timescale for the NCNC process is given by:

$$\tau_{NCNC} \simeq \frac{\eta a_0}{\gamma}\sigma^3 N^{\omega} \qquad \text{(A5)}$$

where ω is in the range 3.8–4.2. There is usually a large uncertainty in the surface tension γ between the hydrophobic residues and water. The range for γ is 25–75 cal Å^{-2} mol^{-1}. The largest timescale to eq. A5 emerges when $\omega \approx 4.2$ and $\gamma \approx 25$ cal Å^{-2} mol^{-1}. Using $\eta \approx 0.01$*Poise*, $\sigma \approx 0.2$, and $a_0 \approx 5 \times 10^{-8}$ cm, τ_{NCNC} ranges from 10^{-6} s to 0.1 ms as N varies from 22 to 66. The values based on theoretical arguments (cf. eq. A5) are consistent with the numerical estimates based on the simulations.

It is interesting to compare the estimates for the fast process, corresponding to the NCNC mechanism, obtained using eqs A4 and A5 and simulation results with experimental results. All the theoretical estimates yield $\tau_{NCNC} \approx 0.1$ ms. The recent experiments on CI2 suggest that the timescale for the nucleation collapse process is in the range 1.5–15 ms (DE Otzen, AR Fersht, personal communication) depending on external conditions. The experimental times are not inconsistent with our simulation results and theoretical estimates given the uncertainty in the values of the various parameters. Our studies further underscore the importance of processes relevant for folding of proteins on a submillisecond timescale especially for the NCNC process. Further experiments on these timescales are needed for an explicit experimental demonstration of the NCNC mechanism [57–59].

Acknowledgements

We are grateful to Alan Fersht for informing us of folding times for CI2 and a mutant of CI2. This work was supported in part by grants from the National Science Foundation (through grant numbers CHE93-07884 and CHE96-29845) and the Air Force Office of Scientific Research. T Veitshans gratefully acknowledges financial support from the Ecole Normale Supérieure de Lyon, France.

References

1. Dill, K.A., *et al.*, & Chan, H.S. (1995). Principles of protein folding – a perspective from simple exact models. *Protein Sci.* **4**, 561–602.
2. Bryngelson, J.D., Onuchic, J.N., Socci, N.D. & Wolynes, P.G. (1995). Funnels, pathways and the energy landscape of protein folding: a synthesis. *Proteins* **21**, 167–195.
3. Wolynes, P.G., Onuchic, J.N. & Thirumalai, D. (1995). Navigating the folding routes. *Science* **267**, 1619–1620.
4. Chan, H.S. & Dill, K.A. (1994). Transition states and folding dynamics of proteins and heteropolymers. *J. Chem. Phys.* **100**, 9238–9257.
5. Thirumalai, D. (1994). Theoretical perspectives on *in vitro* and *in vivo* protein folding. In *Statistical Mechanics, Protein Structure, and Protein Substrate Interactions.* (Doniach, S., ed.), pp. 115–133, Plenum Press, New York.
6. Abkevich, V.I., Gutin, A.M. & Shakhnovich, E. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* **33**, 10026–10036.
7. Socci, N.D. & Onuchic, J.N. (1995). Kinetic and thermodynamic analysis of protein-like heteropolymers: Monte Carlo histogram technique. *J. Chem. Phys.* **103**, 4732–4744.
8. Thirumalai, D. & Woodson, S.A. (1996). Kinetics of folding of proteins and RNA. *Acc. Chem. Res.* **29**, 433–439.
9. Otzen, D.E., Itzhaki, L.S. & Fersht, A.R. (1994). Structure of the transition state for the folding/unfolding of the barley chymotrypsin inhibitor 2 and its implications for mechanisms of protein folding. *Proc. Natl. Acad. Sci. USA* **91**, 10422–10425.
10. Fersht, A.R. (1995). Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proc. Natl. Acad. Sci. USA* **92**, 10869–10873.
11. Radford, S.E. & Dobson, C.M. (1995). Insights into protein folding using physical techniques: studies of lysozyme and alpha-lactalbumin. *Phil. Trans. Roy. Soc. Lond. B* **348**, 17–25.
12. Sosnick, T.R., Mayne, L., Hiller, R. & Englander, S.W. (1994). The barriers in protein folding. *Nat. Struct. Biol.* **1**, 149–156.
13. Schindler, T., Herrler, M., Marahiel, M.A. & Schmid, F.X. (1995). Extremely rapid folding in the absence of intermediates. *Nat. Struct. Biol.* **2**, 663–673.
14. Kiefhaber, T. (1995). Kinetic traps in lysozyme folding. *Proc. Natl. Acad. Sci. USA* **92**, 9029–9033.
15. Itzhaki, L.S., Otzen, D.E. & Fersht, A.R. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analyzed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260–288.
16. Guo, Z. & Thirumalai, D. (1995). Kinetics of protein folding: nucleation mechanism, time scales, and pathways. *Biopolymers* **36**, 83–103.
17. Thirumalai, D. & Guo, Z. (1995). Nucleation mechanism for protein folding and theoretical predictions for hydrogen-exchange labeling experiments. *Biopolymers* **35**, 137–140.
18. Camacho, C.J. & Thirumalai, D. (1995). Theoretical predictions of folding pathways using the proximity rule with applications to BPTI. *Proc. Natl. Acad. Sci. USA* **92**, 1277–1281.
19. Dadlez, M. & Kim, P.S. (1995). A third native one-disulphide intermediate in the folding of bovine pancreatic trypsin inhibitor. *Nat. Struct. Biol.* **2**, 674–679.
20. Mirny, L.A., Abkevich, V. & Shakhnovich, E.I. (1996). Universality and diversity of the protein folding scenarios: a comprehensive analysis with the aid of a lattice model. *Fold. Des.* **1**, 103–116; 1359-0278-001-00103.
21. Thirumalai, D. (1995). From minimal models to real proteins: time

scales for protein folding kinetics. *J. Physique (Paris) I* **5**, 1457–1467.
22. Onuchic, J.N., Wolynes, P.G., Luthey-Schulten, Z.A. & Socci, N.D. (1995). Toward an outline of the topography of a realistic protein-folding funnel. *Proc. Natl. Acad. Sci. USA* **92**, 3626–3630.
23. Klimov, D.K. & Thirumalai, D. (1996). A criterion that determines the foldability of proteins. *Phys. Rev. Lett.* **76**, 4070–4073.
24. Camacho, C.J. & Thirumalai, D. (1993). Kinetics and thermodynamics of folding in model proteins. *Proc. Natl. Acad. Sci. USA* **90**, 6369–6372.
25. Alexander, P., Fahnestock, S., Lee, T., Orban, J. & Bryan, P. (1992). Thermodynamic analysis of the folding of the streptococcal protein G IgG-binding domains B1 and B2: why small proteins tend to have high denaturation temperatures. *Biopolymers* **31**, 3597–3603.
26. Jackson, S.E. & Fersht, A.R. (1991). Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. *Biochemistry* **30**, 10428–10435.
27. Honeycutt, J.D. & Thirumalai, D. (1990). Metastability of the folded states of globular proteins. *Proc. Natl. Acad. Sci. USA* **87**, 3526–3529.
28. Leopold, P.E., Montal, M. & Onuchic, J.N. (1992). Protein folding funnels: a kinetic approach to the sequence–structure relationship. *Proc. Natl. Acad. Sci. USA* **89**, 8721–8725.
29. Honig, B. & Cohen, F.E. (1996). Adding backbone to protein folding: why protein are polypeptides. *Fold. Des.* **1**, R17–R20; 1359-0278-001-R0017.
30. Rey, A. & Skolnick, J. (1991). A comparison of lattice Monte Carlo dynamics and Brownian dynamics folding pathways of α-helical hair-pins. *Chem. Phys.* **158**, 199–219.
31. Garrett, D.G., Kastella, K. & Ferguson, D.M. (1992). New results on protein folding from simulated annealing. *J. Am. Chem. Soc.* **114**, 6555–6556.
32. Guo, Z. & Thirumalai, D. (1996). Kinetics and thermodynamics of folding of a *de novo* designed four-helix bundle protein. *J. Mol. Biol.* **263**, 323–343.
33. Sosnick, T.R., Mayne, L. & Englander, S.W. (1996). Molecular collapse: the rate limiting step in two-state cytochrome C folding. *Proteins* **24**, 413–426.
34. Honeycutt, J.D. & Thirumalai, D. (1992). The nature of folded states of globular proteins. *Biopolymers* **32**, 695–709.
35. Andersen, H.C. (1983). Rattle: a "velocity" version of the shake algorithm for molecular dynamics calculations. *J. Comp. Phys.* **52**, 24–34.
36. Creighton, T.E. (1993). *Proteins: Structures and Molecular Properties.* WH Freeman & Co, New York.
37. McCammon, J.A. & Harvey, S.C. (1988). *Dynamics of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge.
38. Pangali, C., Rao, M. & Berne, B.J. (1979). A Monte Carlo simulation of the hydrophobic interaction. *J. Chem. Phys.* **71**, 2975–2981.
39. Swope, W.C., Andersen, H.C., Berens, P.H. & Wilson, K.R. (1982). A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: application to small water clusters. *J. Chem. Phys.* **76**, 637–649.
40. Amara, P. & Straub, J.E. (1995). Folding model proteins using kinetic and thermodynamic annealing of the classical density distribution. *J. Phys. Chem.* **99**, 14840–14853.
41. De Gennes, P.G. (1979). *Scaling Concept in Polymer Physics.* Cornell University Press, New York.
42. Sali, A., Shakhnovich, E. & Karplus, M. (1994). Kinetics of protein folding: a lattice model study of the requirements for folding to the native state. *J. Mol. Biol.* **235**, 1614–1636.
43. Klimov, D.K. & Thirumalai, D. (1996). Factors governing the foldability of proteins. *Proteins* in press.
44. Straub, J. & Thirumalai, D. (1993). Theoretical probes of conformational fluctuations in S-peptide and RNase A/3′-UMP enzyme product complex. *Proteins* **15**, 360–373.
45. Bryngelson, J.D. & Wolynes, P.G. (1989). Intermediates and barrier crossing in a random energy model (with application to protein folding). *J. Phys. Chem.* **93**, 6902–6915.
46. Go, N. (1983). Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* **12**, 183–210.
47. Shakhnovich, E. & Gutin, A.M. (1993). A new approach to the design of stable proteins. *Protein Eng.* **6**, 793–800.
48. Shakhnovich, E. (1994). Proteins with selected sequences fold into unique native conformation. *Phys. Rev. Lett.* **72**, 3907–3910.
49. Deutsch, J.M. & Kurosky, T. (1996). New algorithm for protein design. *Phys. Rev. Lett.* **76**, 323–326.
50. Camacho, C.J. & Thirumalai, D. (1995). Modeling the role of disulfide bonds in protein folding: entropic barriers and pathways. *Proteins* **22**, 27–40.
51. Camacho, C.J. & Thirumalai, D. (1996). A criterion that determines fast folding of proteins: a model study. *Europys. Lett.* **35**, 627–632.
52. Garel, T., Orland, H. & Thirumalai, D. (1996). Analytical theories of protein folding. In *New Development in Theoretical Studies of Proteins.* (Elber, R., ed.), World Scientific, Singapore.
53. Stillinger, F.H. & Weber, T.A. (1982). Hidden structure in liquids. *Phys. Rev. A* **25**, 978–989.
54. Goldstein, R.A., Luthey-Schulten, Z.A. & Wolynes, P.G. (1992). Optimal protein-folding codes from spin-glass theory. *Proc. Natl. Acad. Sci. USA* **89**, 4918–4922.
55. Jackson, S.E. & Fersht, A.R. (1991). Folding of chymotrypsin inhibitor 2. 2. Influence of proline isomerization on the folding kinetics and thermodynamic characterization of the transition state of folding. *Biochemistry* **30**, 10436–10443.
56. Socci, N.D., Onuchic, J.N. & Wolynes, P.G. (1996). Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.* **104**, 5860–5871.
57. Jones, C.M., *et al.*, & Eaton, W.A. (1993). Fast events in protein folding initiated by nanosecond laser photolysis. *Proc. Natl. Acad. Sci. USA* **90**, 11860–11864.
58. Ballew, R.M., Sabelko, J. & Gruebele, M. (1996). Direct observation of fast protein folding: the initial collapse of apomyoglobin. *Proc. Natl. Acad. Sci. USA* **93**, 5759–5764.
59. Pascher, T., Chesick, J.P., Winkler, J.R. & Gray, H.R. (1996). Protein folding triggered by electron transfer. *Science* **271**, 1558–1560.

Because *Folding & Design* operates a 'Continuous Publication System' for Research Papers, this paper has been published via the internet before being printed. The paper can be accessed from http://biomednet.com/cbiology/fad.htm – for further information, see the explanation on the contents page.