Universality and diversity of the protein folding scenarios: a comprehensive analysis with the aid of a lattice model

Leonid A Mirny, Victor Abkevich and Eugene I Shakhnovich

Background: The role of intermediates in protein folding has been a matter of great controversy. Although it was widely believed that intermediates play a key role in minimizing the search problem associated with the Levinthal paradox, experimental evidence has been accumulating that small proteins fold fast without any detectable intermediates.

Results: We study the thermodynamics and kinetics of folding using a simple lattice model. Two folding sequences obtained by the design procedure exhibit different folding scenarios. The first sequence folds fast to the native state and does not exhibit any populated intermediates during folding. In contrast, the second sequence folds much slower, often being trapped in misfolded low-energy conformations. However, a small fraction of folding molecules for the second sequence fold on a fast track avoiding misfolded traps. In equilibrium at the same temperature the second sequence has a highly populated intermediate with structure similar to that of the kinetics intermediate.

Conclusions: Our analysis suggests that intermediates may often destabilize native conformations and derail the folding process leading it to traps. Less-optimized sequences fold via parallel pathways involving misfolded intermediates. A better designed sequence is more stable in the native state and folds fast without intermediates in a two-state process.

Address: Harvard University, Department of Chemistry, 12 Oxford Street, Cambridge, MA 02138, USA.

Correspondence to: Eugene I Shakhnovich
E-mail address: eugene@diamond.harvard.edu

Key words: folding intermediates, lattice model, molten globule, protein folding

Received: **08 Dec 1995**
Accepted: **02 Jan 1996**

Published: **20 Feb 1996**
Electronic identifier: **1359-0278-001-00103**

Folding & Design 20 Feb 1996, **1**:103–116

Introduction

For a long time, the protein folding field was dominated by the concept of intermediates as necessary elements for solution of the Levinthal paradox [1]. The search for intermediates in protein folding was motivated by the assertion that hierarchical and parallel elements could significantly simplify the search process. The models included framework [2–3] and diffusion-collision [4] where a hierarchical mechanism was postulated, i.e. secondary structure elements forming first, getting 'frozen' and subsequently moving as a whole, thus decreasing the number of degrees of freedom to be searched. The key point of these models was that folding starts from small-scale details (i.e. at the level of local contacts in the sequence), and elements of correct structure progressively propagate on larger scales until the whole molecule is folded.

Such a hierarchical folding scenario possesses certain features of self-similarity. The same scenario is known in physics to take place at the second-order phase transitions at which a physical system does not overcome any free energy barriers and there is no rate-limiting stage [5–7]. A possible example of such a self-similar process is collapse of a homopolymer [8] resulting in a 'crumpled' globule in which local (along the sequence) contacts dominate [9]. The concept of intermediates fits naturally into the hierarchical (second-order-like) scheme, as it suggests that structural features of the native conformation evolve in time sequentially propagating from small-scale details to full structure.

An alternative scenario is a 'first-order-like' folding which entails overcoming the major free energy barrier. The rate-limiting step for this process is the transition from the free energy minimum, corresponding to the unfolded state (or burst intermediate, see below) to the transition state for folding. Subsequent descent to the native state is fast. In contrast to the hierarchical scenario, the transition state scenario is not self-similar, as at least two kinetic steps can be distinguished in each folding event: going 'uphill' to the transition state then rapid descent to the folded state from the transition state.

In the first-order-type transition, the two phases (in the case of proteins folded and unfolded) are separated by the free energy barrier and they do not usually have very much in common. In this case, the existence of certain elements of native conformation in the unfolded state may stabilize the latter, i.e. increase the kinetic barrier and decrease the thermodynamic stability [10]. The rate of folding may be significantly decreased if contacts present in the unfolded state are not the ones formed in the transition state. These can be either non-native contacts or native-like contacts that get consolidated only upon final

folding from the transition state to the native state [10–12]. Such contacts must first be broken to make it possible for the chain to form the transition state. This factor decelerates folding.

The thermodynamic analysis [13,14] suggests that the folding transition is cooperative and first-order-like (as for the definition of 'phase transitions' in such finite systems as biological macromolecules [7,15]).

Experimentally, kinetic intermediates were indeed found for some proteins [16–18]. However, recent studies for a number of proteins (mostly small ones without disulfides) showed that fast folding can take place without intermediates [19–25]. Not surprisingly, the existing opinions about the role of intermediates in protein folding are as diverse as the folding scenarios themselves. They range from the assertion that intermediates are always necessary to resolve kinetic problems of folding [26,27] to the point of view that intermediates may be 'adventitious', and their presence may even slow down folding [10,23,28].

To resolve this critical issue, both theoretical and experimental approaches should be applied. Experimental approaches to protein folding, despite their importance, have certain limitations. These include serious difficulties in structural characterization of intermediates and transition states, especially at the fast stages of folding. More importantly, all experimentally measured quantities are averaged over the ensemble of protein molecules, whereas in simulation folding each molecule can be followed and analyzed.

It is important to note that the number of scenarios is quite limited (much less than the number of proteins) which suggests that there are generic features of the folding process that are universal for many proteins. This property lends itself naturally to theoretical analysis which, when successful, provides a general point of view and distinguishes between generic features of the phenomena it studies and its consequences, which may or may not occur.

The important requirement of the theoretical models of protein folding is that they reproduce the folding phenomenon in a totally unbiased calculation. Simulations, starting from random coils, in all runs should converge to a unique folded conformation which is stable under a wide variety of conditions. The calculations must be efficient, i.e. allow for hundreds of runs to make all conclusions statistically grounded.

The only available theoretical systems in which these requirements can be met are lattice models [29–37]. Their numerically exact character and the possibility of addressing eventually any meaningful question make them a valuable tool for understanding the mechanism of protein folding.

In this work, we study the kinetics and thermodynamics of two particular sequences obtained by the design procedure. These sequences (termed here *Seq1* and *Seq2*) were chosen for comparative analysis because they have different thermodynamic behavior. Although both of them fold to the same native state, *Seq2* (in contrast to *Seq1*) has an equilibrium intermediate.

We show that these two sequences follow different folding kinetics. Particularly, *Seq2* is often trapped in misfolded conformations on its way to the native state. We characterize the structures of these misfolded intermediates and show that they are responsible for slowing down the folding kinetics of *Seq2*. By studying folding trajectories of a single molecule, we show that a small fraction of *Seq2* molecules fold on a fast track to the native state avoiding misfolded traps. In contrast, no kinetic intermediates were detected for *Seq1*, which folds much faster than *Seq2*. In equilibrium, a significant fraction of *Seq2* molecules is folded into a compact non-native conformation. This equilibrium intermediate has a structure similar to those of kinetic intermediates. We identify the non-native contacts in *Seq2* that are responsible for misfolding to kinetic traps, conformations which also constitute populated equilibrium intermediates.

Model

We consider the conformation of a protein chain as a self-avoiding walk on a cubic lattice. The energy of a conformation is the sum of energies of pairwise contacts between monomers which are not nearest neighbors in sequence:

$$E = \sum_{1 \le i < j \le N} U(\xi_i, \xi_j)\Delta_{ij} \qquad (1)$$

where $\Delta_{ij}=1$ if monomers i and j are lattice neighbors and $\Delta_{ij}=0$ otherwise. ξ_i defines the type of amino acid residue in position i. $U(\xi,\eta)$ is a magnitude of contact interaction between amino acids of types ξ and η (taken from Table 6 in [38]).

The most important step is the selection of sequences that fold to their native state and are stable in that conformation.

It has been shown [12,34,35,39–41] that the necessary and sufficient condition for sequences to fold in this model is for the native state to be a pronounced energy minimum for this sequence, compared to the set of misfolded conformations. Hence, the sequence design was aimed at generation of such sequences. The detailed discussion of the design algorithm, which is Monte Carlo optimization in sequence space, is published in our previous works [12,42,43].

The procedure is as follows: first, choose an arbitrary conformation of the lattice protein chain to serve as the native

structure; second, select sequences that have low energy in this 'native' structure compared to unfolded and misfolded conformations; and third, fold the designed sequences (from the previous step) using Monte Carlo simulations.

An important feature of this approach is that it provides many non-homologous sequences that fold to the same native conformation. Analysis of their folding behavior makes it possible to determine which features of the folding scenario are generic, which are due to the character of interactions in a given sequence, and which are due to structural features of the native conformation [44].

The design approach requires first the selection of a target conformation for which sequences should be designed so that this target conformation will be 'native' for them (i.e. stable and kinetically accessible). We used the conformation shown in Figure 1 as the target; this 'native' conformation is the same as was used in our previous studies [12,40].

Somewhat different design procedures were used to generate *Seq1* and *Seq2*. *Seq2* was generated using the approach described in detail in our previous papers [42,43]. This is the optimization of energy of the native conformation with respect to permutation of sequences with fixed amino acid composition. (We used the composition corresponding to the 'average' protein [45]; *Seq2* is the same as used in [12].)

Seq1, on the other hand, was designed using an improved design method. In this variation on our method, what we minimize is not the energy of the native conformation E_{nat} (as for *Seq2*), but its relative value [46,47]:

$$E_{rel} = \frac{E_{nat} - E_{av}}{\sigma} \qquad (2)$$

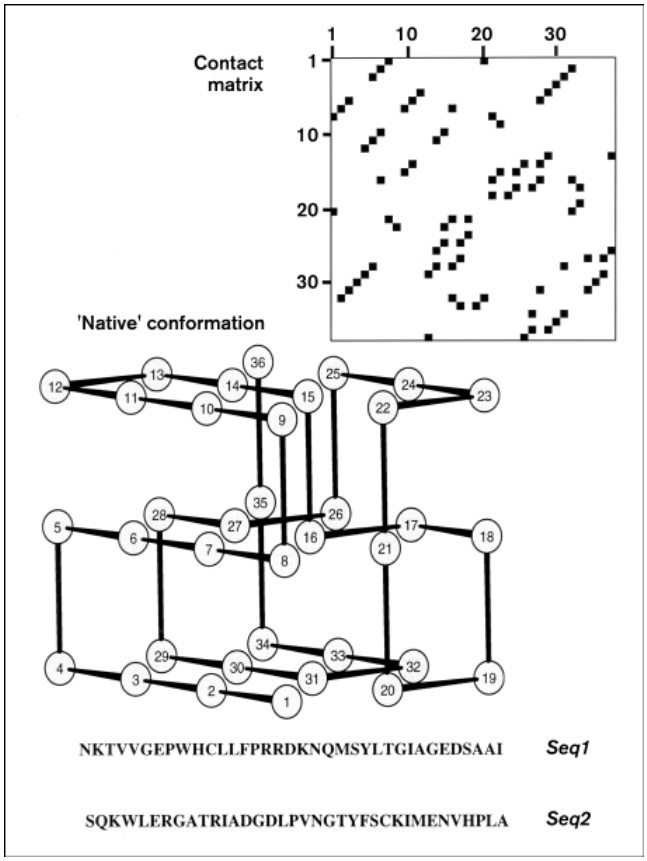
where E_{av} is the average energy of non-native conformations and σ is the standard variance of energies of all contacts. To estimate their values, we first compute the energies of all topologically possible contacts between all monomer pairs. From this, we calculate the average energy e_{av} of a contact and standard variance σ and then estimate the average energy of non-native conformation as $E_{av}=N_{total} \times e_{av}$, where N_{total} is the total number of contacts in the native conformation.

Note that *Seq1* and *Seq2* have the same native conformation (Fig. 1). We made several long runs in order to make sure that no conformations with energy lower than that of the native state are encountered. This was indeed the case, which made us believe that the target conformation is the global energy minimum for each of the these two sequences.

Folding simulations were carried out using the standard Monte Carlo method for polymers on a cubic lattice. (For a detailed discussion of lattice Monte Carlo simulation technique, its advantages and caveats, see [29,34,48].) Our analysis represents a simulation counterpart of stopped-flow protein folding experiment. Each simulation started from a random coil conformation, proceeded at constant temperature and lasted about five times longer than the mean folding time. This corresponds to a stopped-flow experiment, in which protein solution is rapidly diluted from, say, 6 M guanidium hydrochloride (GdmHCl) (or low pH) where chains are in or near the random coil state to a lower concentration of GdmHCl (or higher pH) where native conformation is thermodynamically stable.

In our modeling, we average over the ensemble of molecules by making 100 runs at each temperature. Different parameters of the folding molecule are calculated at every 1000 Monte Carlo (MC) step. We then compute average values of all parameters sampled during 40 000 MC steps. This sampling technique models an experimental pulse-

Figure 1

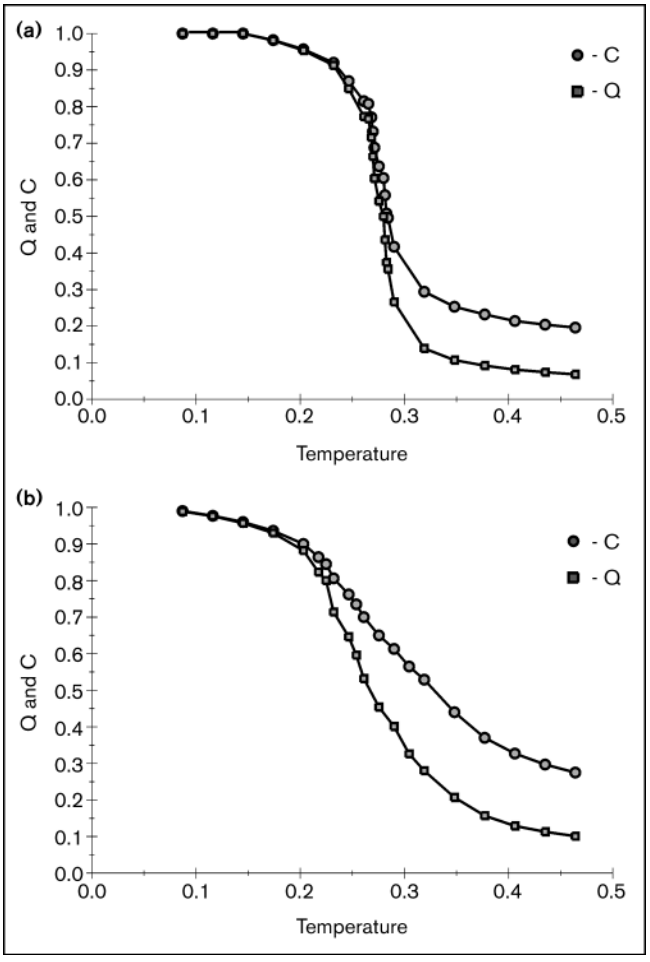


The 'native' conformation and two sequences, *Seq1* and *Seq2*, designed to fold into this conformation. In the contact matrix, a position (*i*,*j*) is marked with a black dot if residues *i* and *j* are non-covalent neighbors in the native structure below. For example, position (16,27) on the contact matrix is marked by a black square because residues 16 and 27 are in contact in the native structure.

labeling procedure in which measurements take a relatively short time interval. On the other hand, sampling performed in this way is equivalent to the moving window average technique, which is widely used in time series analysis to filter the time series out of high frequency noise.

A number of characteristics are of interest to us. One is the compactness characterized by the normalized number of contacts in conformations, averaged over all runs. $C=N/N_{total}$ where N is the number of all contacts in a conformation, N_{total} is the number of contacts in a maximally compact conformation; N_{total}=40 for a 36-mer. If C=1, a conformation is maximally compact. A second characteristic is the degree of folding, defined as $Q=N_{native}/N_{total}$ where N_{native} is the number of native contacts in a conformation. Q=1 in the native conformation. A third characteristic is the occupancy of all individual native contacts at MC time t, $f_{ij}(t)$. This quantity is defined in a straightforward way. Looking at 40 conformations (which were sampled during the time interval [t,t+40 000] MC steps), we count the fraction of conformations that have a contact between residues i and j.

Figure 2



Equilibrium compactness C and degree of folding Q as a function of temperature for **(a)** *Seq1* and **(b)** *Seq2*. These quantities were obtained at each temperature by averaging compactness and degree of folding over 10 MC runs each 10^7 steps.

Results

Thermodynamics

In order to study the thermodynamics for the two sequences (*Seq1* and *Seq2*), we performed MC folding simulations at different temperatures. The equilibrium values of C and Q were computed by averaging over uncorrelated conformations obtained from several MC runs (see Materials and methods for details). The temperature dependence of the equilibrium C and Q values is shown in Figure 2. Both sequences exhibit folding transition to the compact native state, which is seen as a rapid increase of C and Q values as temperature decreases.

The character of this transition is different, however, for these two sequences. The first sequence, *Seq1*, exhibits a single transition both in C and Q, as both curves go in a coherent way (Fig. 2a).

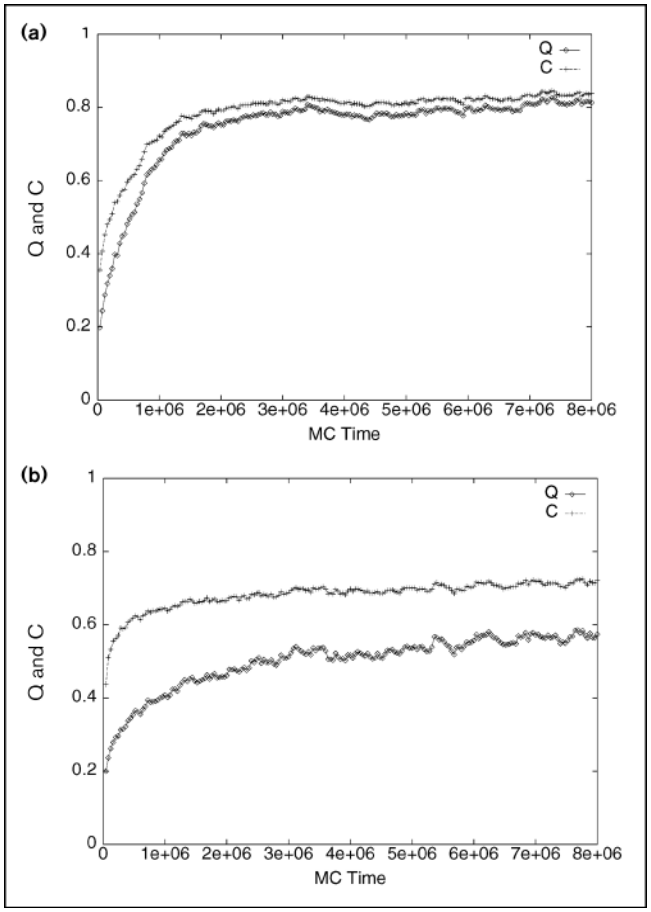
In contrast, the second sequence, *Seq2*, first becomes compact and non-native (high C and low Q) as temperature decreases and then, at lower temperatures, it folds to the native conformation (Fig. 2b).

This character of equilibrium transition in *Seq2* is very important. From Figure 2b, one can see that there is a temperature range at which *Seq2* is relatively compact and has a non-native conformation. Note that even at these temperatures, *Seq2* folds to its native conformation. What are these non-native compact conformations which dominate in the ensemble at these temperatures? How does this sequence fold into its native conformation? To address the question of how the sequences fold, we turn to kinetic simulations of *Seq1* and *Seq2*.

Kinetics

Simulations of folding kinetics were performed at T=0.26. At this temperature in equilibrium *Seq1* is in the stable native conformation (high C and high Q) whereas *Seq2* is in compact, but not highly native conformation (high C and lower Q). For each sequence, we performed 100 MC runs starting from random initial conformations. Different runs simulate independently folding molecules. By averaging over several runs, we model the ensemble of folding molecules. Data obtained in this way can be compared with those obtained in experiments. The advantage of the simulation is that we can also study each folding trajectory separately, revealing basic events in the protein folding reaction of each molecule.

Figure 3



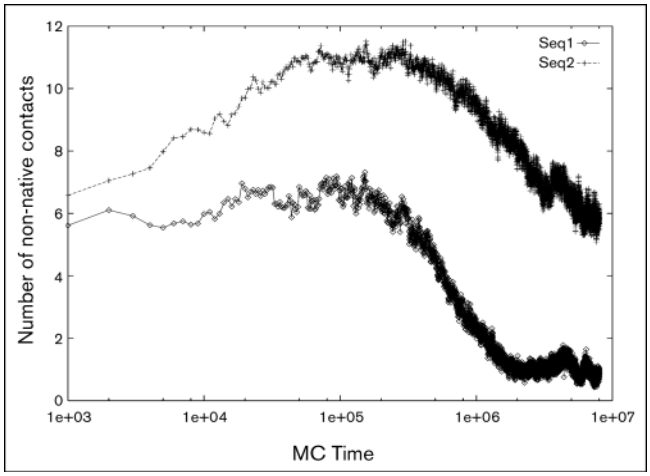
The MC-time dependence of average (over 100 runs) compactness and degree of folding for **(a)** *Seq1* and **(b)** *Seq2*.

The following results were obtained for *Seq1* and *Seq2* in 100 MC runs. Firstly, *Seq1* exhibited better 'foldicity' by folding to the native conformation in all 100 runs, whereas *Seq2* reached the native state during the time of simulation (8 000 000 MC step) in only 84 runs.

Secondly, *Seq1* has an overall rate of folding higher than that of *Seq2*. First passage time (FPT) is measured as the number of MC steps made by a molecule in one run to reach the native state. *Seq1* has a much narrower distribution of FPT than *Seq2*. Such a wide dispersion of *Seq2* folding times is suggestive of different runs following principally different trajectories of folding. We discuss this in more detail below.

Thirdly, we also found that during folding *Seq2* spends a long time in compact (high C) and non-native (low Q) conformations. This behavior is not characteristic for *Seq1*. These findings were based on the analysis of average compactness C and degree of folding Q plotted as a function of MC time (see Figs 3,4).

Figure 4



The MC-time dependence of average (over 100 runs) number of non-native contacts ($C-Q$) for both sequences. The maximum on the curve corresponding to *Seq2* is the signature of the kinetic intermediate for this sequence.

The conclusion to be drawn from these observations is that *Seq1* and *Seq2* have different folding scenarios which should be studied in detail.

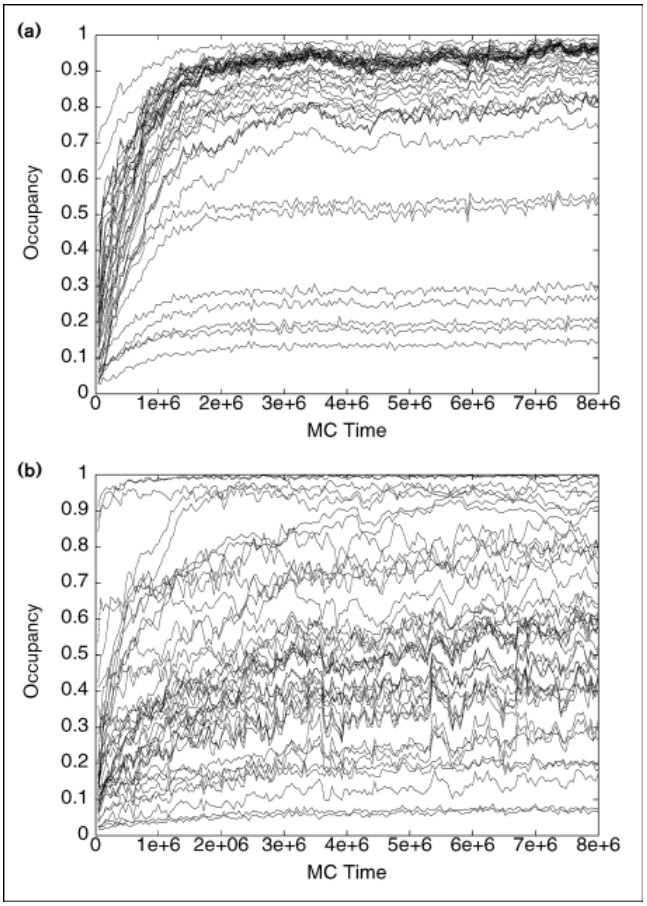
Contact kinetics

In order to reveal the difference between the folding kinetics of *Seq1* and *Seq2*, we considered the kinetics of a single contact formation.

The method we applied to study MC trajectories is similar in spirit to the NMR pulse labeling technique used in protein folding experiments [49–51]. Namely, in every 40 000 MC steps we made snapshots of the folding molecule. During the snapshot, 40 structures were sampled with the time interval of 1000 MC steps and then the average structure was recorded. Similar to the NMR technique, the average structure is represented by occupancies of all residue–residue contacts in the sampled structures. For example, if in the average structure the occupancy of a contact between amino acids 1 and 4 equals 0.8, then this contact was present in 32 out of 40 structures sampled during the interval of 40 000 MC steps. Hence, the snapshot is the average contact map of conformations present during a short interval of folding trajectory.

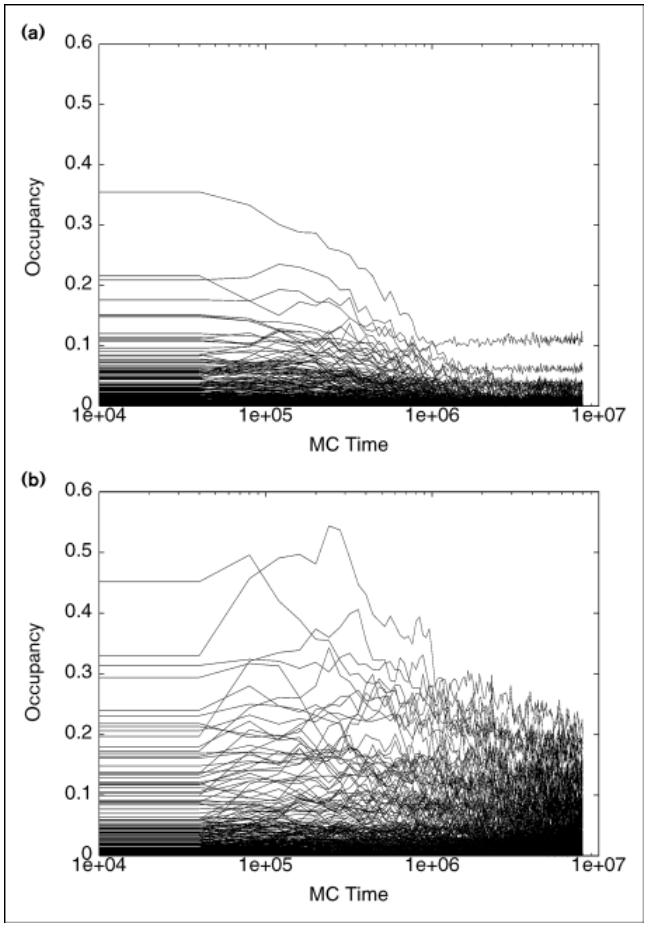
The set of snapshots collected for each run comprises a complete folding trajectory. Averaging over the snapshots obtained at the same time from different runs allows us to model over the population of different molecules in the experiment. As for NMR experiments, the fraction of built native contacts was computed as a function of time as well as occupancy of each contact (native or non-native) as a function of time. Figure 5 presents the occupancy of

Figure 5



Kinetics of formation of individual native contacts for both sequences. **(a)** Native contacts for *Seq1*. **(b)** Native contacts for *Seq2*. These data were obtained from analysis of snapshots taken every 1000 steps in 100 independent runs. The fractions of conformations (out of 100) having specific contact are plotted as kinetic curves. Each curve corresponds to a specific contact (40 curves total). The plots illustrate wide dispersion of rates and final amplitudes for *Seq2* in contrast to much less dispersion for *Seq1*.

Figure 6



Kinetics of formation of individual non-native contacts for both sequences. **(a)** Non-native contacts for *Seq1*. **(b)** Non-native contacts for *Seq2*. The log scale for MC-time axis was used to emphasize fast stages at which the intermediate is formed. The procedure for obtaining the data for these curves is the same as described in the caption to the previous figure. Kinetics for each non-native contact are shown (total 249 curves on each plot). A number of non-native contacts become significantly populated in the kinetic intermediate of *Seq2* (at MC-time ~10^5).

each native contact as a function of folding time for *Seq1* and *Seq2*. The same information for non-native contacts is given in Figure 6.

The analysis of the data for single-contact kinetics leads to the following conclusions.

Native contacts

Firstly, contacts formed by amino acids located near each other along the chain (local contacts) are formed much faster than other contacts. Strong attractive local contacts are formed in the first 10 000–40 000 steps and become extremely stable with an occupancy of 0.9–0.95 for the whole duration of simulation. The rate of folding is at least one order of magnitude higher for local strong contacts than the total rate of folding.

Secondly, contacts formed by amino acids located at the ends of the chain are less stable than contacts formed by amino acids located in the middle of the chain. One can see from Figure 5 that for both sequences there are native contacts that have relatively low occupancy in the folded state. These less stable contacts are formed by the amino acids located at the end of the chain. The reason for such low occupancy of these contacts is the high flexibility of the chain ends.

Thirdly, most native contacts in *Seq1* (excluding the local contacts and the end) have the same rate of folding and the same equilibrium occupancy (see Fig. 5). Hence, the folding process goes in one step for *Seq1* with the single

characteristic rate. For *Seq2*, by contrast, there is almost no characteristic rate of contact formation because the distribution of folding rates is very broad and in fact for each contact the folding kinetics are not exponential (see Fig. 7). All contacts in *Seq2* have different rates, providing no one-step folding transition for the whole chain (see the example of a single contact kinetics in Fig. 7). Attempts to describe contact kinetics for *Seq2* by a double-exponential model resulted in much better fits (Fig. 7). The issue arises as to whether this is simply due to the fact that more fitting parameters were used or double-exponential kinetics reflect important features of folding for *Seq2*. We note that the slow rate constant varies from contact to contact for *Seq2* much less than the fast rate constant. The latter varies from contact to contact in a wide range, and therefore fast constants are not likely to describe any cooperative relaxation characteristic of the whole molecule.

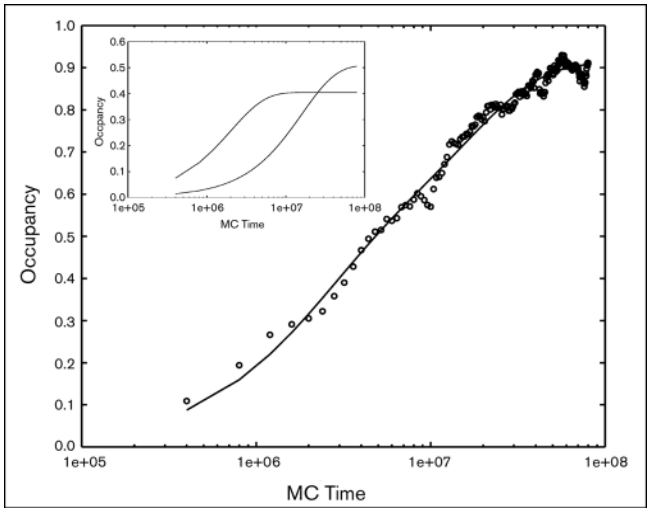
The physical picture consistent with these observations is that the intermediate undergoes continuous relaxation, which cannot be characterized by any particular time constant. However, transition from this intermediate to the native state is a cooperative rate-limiting process in which the whole molecule undergoes the transition over the main (nucleation [10,12,52,53]) barrier to the native state.

Non-native contacts

Occupancy as a function of MC time is shown for all non-native contacts in Figure 6. We use log scale for MC time to emphasize the first stages of folding where non-native interactions might play an essential role in trapping the folding process. As one might expect, the occupancy of non-native contacts decays with time. This decay is different for *Seq1* and *Seq2*, however. For *Seq1*, the occupancy of non-native contacts decays fast with time and soon reaches the marginal level f=0.05–0.1. For *Seq2*, by contrast, significant growth of occupancy for some non-native contacts precedes the decay. Some contacts reach a very high level of occupancy at intermediate time f=0.5. This increase in occupancy of non-native contacts is observed for the time interval of 100 000–500 000 MC steps. No non-native contacts with high occupancies were observed for *Seq1*, which folds much faster than *Seq2* (see above). To check whether non-native contacts which appear at the early stages of folding are responsible for misfolding of *Seq2* and, hence, slowing down its folding reaction, we examined the snapshots obtained at different stages of the folding process.

To study the role of non-native contacts in folding kinetics, we focus on the interval of 100 000–500 000 MC steps, when the average occupancy of non-native contacts reaches its maximum. There are few contacts with high occupancies that dominate at this time interval in all runs. To gain further insight into the structural role of these dominating non-native contacts, we built a snapshot corresponding to the average over all chain conformations found in the interval of interest in all runs (see Fig. 8). Contacts with high occupancy (shown in dark gray) characterize mostly populated conformations at this time interval. These dominating contacts can be categorized into three groups according to their structural role: local contacts, which correspond to turns and helixes (contacts: 3–6; 7–10; 11–14; 27–30); the first anti-parallel sheet, connecting the ends of the chain (contacts: 3–30; 4–29; 5–28; 6–27); and the second anti-parallel sheet, terminated by a turn (contacts: 3–14; 4–13; 5–12; 6–11; 7–10).

Figure 7

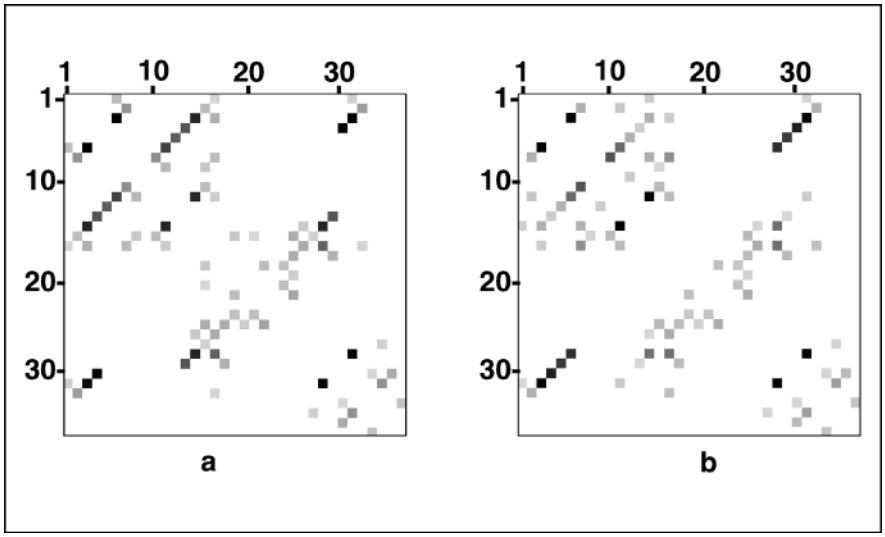


An example of contact kinetics for one of the contacts (5–28) in *Seq2*. Insert shows separately the fast and slow exponential components.

All local contacts and the first sheet contain only the native contacts, whereas the second sheet contains two non-native contacts (3–14; 4–13). These contacts have very high occupancy f_{3-14}=0.72; f_{4-13}=0.75 at the time interval of 100 000–300 000 MC steps. The absence of a clear pattern of contacts in the other regions in the snapshot demonstrates that a large part of the chain is still disordered and the only structural elements present at this time interval are those described above. Information contained in the snapshot is sufficient to reconstruct the topology of this intermediate.

The contact matrix shown in Figure 8 represents schematically the structure of this intermediate. Comparison of intermediate topology (Fig. 9) with the conformation of the native state (Fig. 1) reveals the basic structural differences. Although only two non-native contacts appear in the intermediate, the overall topology of the intermediate is very different from the native topology. The sheet containing non-native contacts is quite stable and prevents folding of several regions of the native structure. Namely, formation

Figure 8



(a) The contact matrix of the kinetic intermediate of *Seq2*. It was obtained by averaging the contact maps of conformations recorded in the 100 000–300 000 MC step interval (every 1000 steps). Dots of different grayscale density denote populations of contact in the kinetic intermediate. The darker the dot, the more often a contact was found in conformations of the kinetic intermediate. **(b)** Contact matrix of the equilibrium intermediate of *Seq2* at T=0.25. This matrix was obtained by averaging contact matrices of the individual conformations having an intermediate degree of folding $0.35<Q<0.55$. These conformations correspond to the maximum of the probability distribution of different values of Q observed in long simulations (of 8×10^7 steps). Conformations were recorded every 1000 steps. Both contact matrices (a) and (b) yield the same chain topology, as shown in Fig. 9.

of non-native sheet 3–14; 4–13 is due to 'incorrect placement' of a large chain fragment (amino acids 9–15) and consequently prevents the correct folding and formation of the following native contacts 13–36; 14–25; 15–24; 15–22 and 9–22. Such misfolding also promotes formation of the two non-native contacts 3–16 and 6–11, which also have high occupancy (see Fig. 8) in this intermediate.

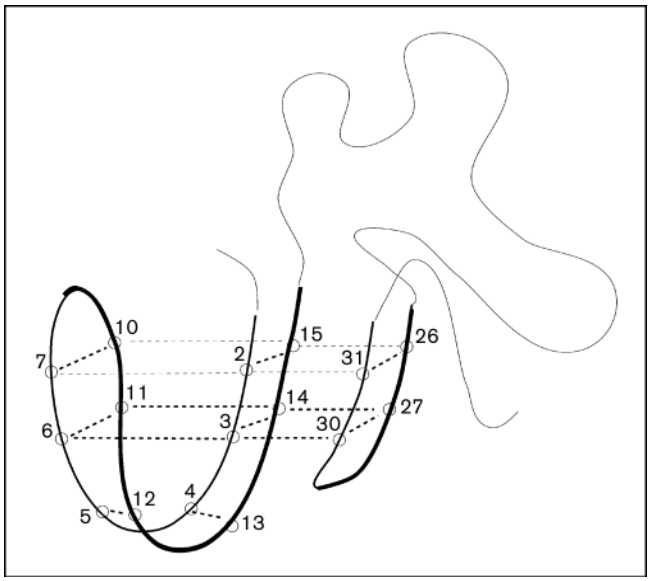
For comparison, we constructed a snapshot for the same time interval for the fast folding sequence *Seq1* (not shown). The pattern of contacts in this snapshot is similar to those of the native state. Almost all native contacts are formed by the time of 100 000–300 000 MC steps and no non-native contacts are found to have high occupancy. Hence, the population of *Seq1* chains has native-like structure at this time.

From the study of an average snapshot at intermediate times for *Seq2*, we conclude that high values of C at early stages of *Seq2* folding are caused by formation of a set of specific non-native contacts which stabilize a misfolded long-living intermediate. Such an intermediate traps the molecule on its folding pathway (Fig. 9).

Note that the misfolded conformation described above is not the only kinetic trap for *Seq2*. A more detailed study of folding trajectories reveals a number of different misfolded traps in which the chain spends a relatively long time (~10^5 MC steps, ~10% of average folding time). Moreover, escaping from one trap, the chain can be captured by another (or the same one). Although the conformations of kinetic traps are highly diverse, they all have a very similar set of non-native contacts stabilizing the misfolded conformations. Namely, contacts 3–14, 3–16, 6–11, 21–24, 7–14 and 32–35 always have high occupancy in the kinetic traps of *Seq2*. This observation is important, as it will be shown below that such contacts are also present in the equilibrium intermediate of *Seq2*.

This examination of the snapshot is close in spirit to NMR pulse-labeling experiments, where one can monitor changes in hydrogen protection with time and make conclusions about the structures of intermediate conformations in the population of folding molecules. We can now take advantage of computer simulations and study the folding trajectory of a single molecule.

Figure 9



The topology of one of the misfolded traps for *Seq2*.

In each run, we selected intervals of folding trajectories in which the molecule had more than 10 non-native contacts. These intervals were assumed to correspond to the misfolded conformations of the chain. If, in addition, the molecule spends a significant amount of time (>200 000) in this state, we consider this state as a kinetic trap. For each trap, we constructed a snapshot which contains average occupancies of all contacts present at this trap. These snapshots obtained for different runs differ significantly from each other. Hence, there are several different misfolded conformations that can trap a molecule for a significant amount of time.

We also compared the snapshots obtained from different runs for the same time interval and observed dramatic differences in patterns of contacts which have high occupancy. In different runs, the chain was found to be either trapped in the misfolded state or be folded to the native state.

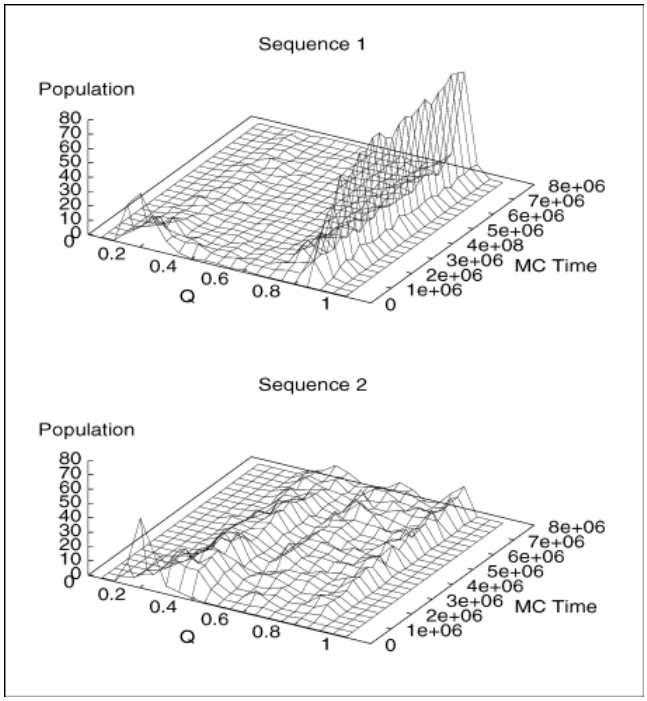
These examples bring us to the conclusion that for *Seq2* different molecules follow distinct folding trajectories and may be either trapped in different misfolded conformations or fold on a fast track to the native state. This heterogeneity in folding trajectories gives rise to a high dispersion of folding times as well as the double exponential kinetics of folding and contact formation described above. This scenario leads to the existence of two population of molecules: those trapped with intermediate values of Q (0.3–0.5), and those folded with high Q (>0.9). To check the conclusion that different molecules follow distinct folding trajectories, we computed the distribution of Q values for the population of molecules at various times. For comparison, this procedure was done for *Seq2*, which has a misfolded intermediate on the folding pathway, and for *Seq1*, which folds faster and does not exhibit any intermediates on the folding pathway.

Figure 10 represents histograms of Q value obtained at different time intervals for the population of 100 folding molecules (i.e. 100 runs). The dramatic difference in folding kinetics of *Seq1* and *Seq2* is easily seen from these plots. There are two humps on the Q distribution for *Seq1* representing folded (Q=0.8–0.9) and unfolded (Q=0.15–0.25) states. Note that only one of these states is populated at any time. As folding starts, the molecules pass from the unfolded state to the folded state, which is seen as a fast decrease of the hump at low Q and an increase of those at high Q. States corresponding to the intermediate values of Q stay almost non-populated, which means that *Seq1* has no long-living intermediates on its folding pathway. In contrast, *Seq2* has three distinct states corresponding to misfolded intermediates (Q=0–0.45), the native state (Q=0.8–0.9) and the native domain state (Q=0.55–0.65). The state with $Q \approx 0.6$ should be identified as a domain state rather than a trap because the kinetics of transitions between this state and the native state are very fast. The existence of the native-like domain has been discussed earlier (see [54]). The other two states, intermediate and folded, are relevant for our discussion. As folding starts, both native and intermediate states become populated. In contrast to *Seq1*, both states stay highly populated during the whole time of simulation. This suggests that the molecules are interconverting between the native and the intermediate states. Folding transition does not have the two-state character for *Seq2*. The analysis of the Q distributions as a function of time reveals the basic difference between the folding kinetics of *Seq1* and *Seq2*. *Seq2* molecules follow parallel pathways and spend some time trapped in a different misfolded conformation. In contrast, *Seq1* folds much faster, exhibiting no long-living intermediates or kinetic traps.

Thermodynamics versus kinetics

In order to study how the kinetic behaviors of the sequences are related to their thermodynamic properties, we performed equilibrium simulations for *Seq1* and *Seq2*. Particularly, we address the issue of whether non-native contacts, which are responsible for trapping the molecule during folding, are present in equilibrium. By making MC runs 10 times longer than those in kinetic simulations (80 000 000 MC steps), we reached the equilibrium conditions for the molecule and computed the equilibrium occupancy of all contacts. Again, among non-native con-

Figure 10



Time evolution of distribution between species with different degrees of folding for *Seq1* and *Seq2*. The statistics are taken over 100 runs.

tacts there are a small set of dominating contacts with high occupancy. Note that non-native contacts with high equilibrium occupancy are the same as non-native contacts responsible for trapping the folding trajectory. Hence, the equilibrium intermediate has a fold similar to the fold of kinetic intermediates. The snapshot of the equilibrium intermediate was obtained by selecting compact partly folded conformations (intermediate values of $0.35<Q<0.55$ and high values of C) from the equilibrium ensemble (Fig. 8b). This procedure corresponds to the selection of conformation belonging to the minimum of free energy corresponding to the intermediate.

This equilibrium intermediate coexists with the folded state in the equilibrium ensemble. The same procedure was performed for *Seq1*. Non-native contacts have negligible occupancy (<0.1) and the general pattern of contacts is the same as in the native state. No equilibrium intermediate was observed for *Seq1*.

To study the temperature dependence of contact occupancies, we performed long MC simulations at T=0.26 to determine the density of states. Specifically, we determined $\nu_{CON}(E)$, the frequency with which states with energy E having a specified contact CON were found in a long simulation run.

Having determined this value, the contact occupancy f_{CON} as a function of temperature is determined for each contact using the histogram technique [55–57]:

$$f_{CON} = \frac{\int dE\, \nu_{CON}(E)\exp(-E/kT)}{\sum_{CON}\int dE\, \nu_{CON}(E)\exp(-E/kT)} \qquad (3)$$

The results of this procedure are shown in Figure 11 (native contacts) and Figure 12 (non-native contacts) for *Seq1* and *Seq2*. Note the important difference in thermodynamic behavior for *Seq1* and *Seq2*.

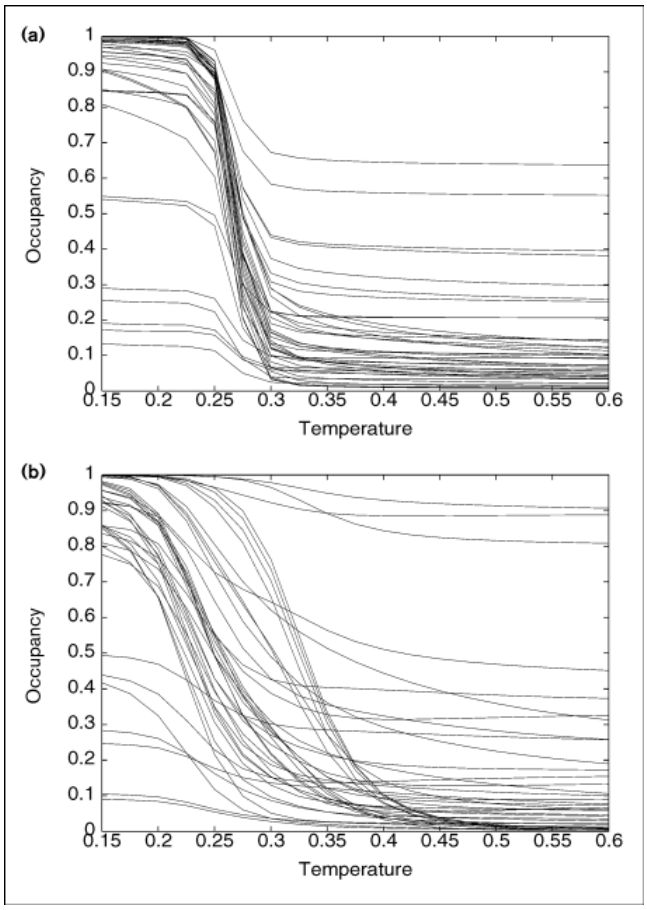
Native contacts

Almost all native contacts of *Seq1* exhibit transition to the native state at the same temperature, whereas native contacts of *Seq2* have different transition temperatures. Transition takes place in the narrow temperature interval for *Seq1* and in the wide temperature interval for *Seq2*. Hence, as temperature decreases, *Seq1* undergoes much more cooperative transition to the folded state than *Seq2*. For *Seq2*, one can roughly distinguish two transitions: a group of weaker contacts decaying at $T\approx0.25$ and another set of stronger contacts which persist up to T=0.35. It is clear that at $0.25<T<0.35$, intermediate 'molten-globule-like' conformations are stable which have few native contacts as well as few non-native contacts (see below).

Non-native contacts

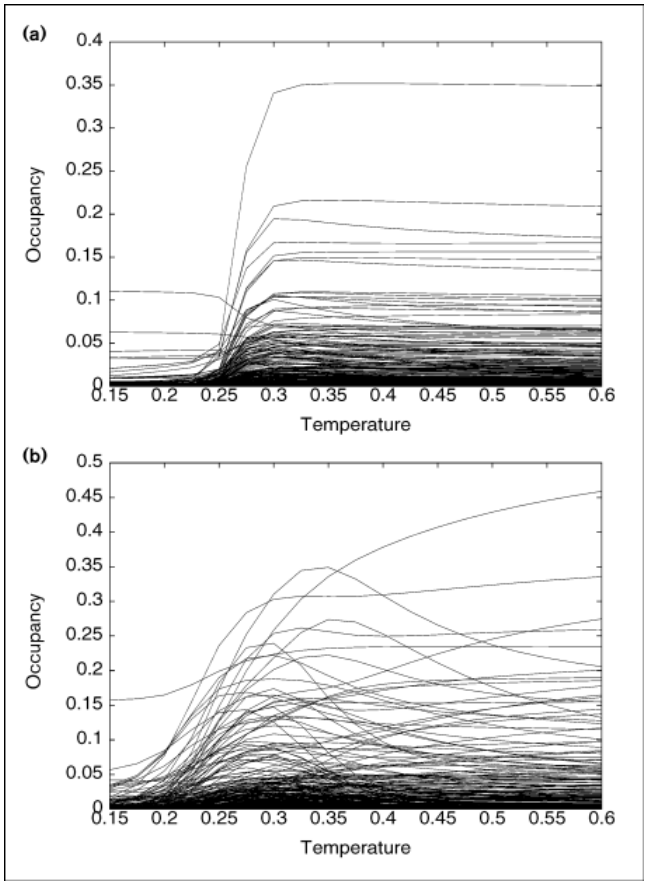
For both sequences, almost all non-native contacts have negligible occupancy at $T<0.2$. As temperature increases, however, non-native contacts of *Seq1* and *Seq2* exhibit different behavior. Non-native contacts of *Seq1* exhibit sharp transition to the level corresponding to the unfolded state (higher occupancy) at T=0.27 and stay at a constant level at high temperatures. In contrast, several non-native contacts of *Seq2* have a maximum at the intermediate temperatures $0.25<T<0.45$ and reach a value 0.2–0.35 of equilibrium occupancy. Hence, in this interval of temperatures an equilibrium misfolded intermediate coexists with the native state for *Seq2*. As temperature increases or decreases the intermediate is destabilized and either the folded state (at low temperatures) or the unfolded state (at high temperatures) become dominantly populated at equilibrium.

Figure 11



The temperature dependence of equilibrium occupancy of all native contacts in **(a)** *Seq1* and in **(b)** *Seq2*. An equilibrium intermediate having some native and some non-native contacts can be distinguished for *Seq2* at $0.27<T<0.33$. Each curve was calculated using the histogram method described in the text. Local contacts have high equilibrium occupancy even at high temperature.

Figure 12



The temperature dependence of equilibrium occupancy of all non-native contacts of **(a)** *Seq1* and **(b)** *Seq2*. The non-native contacts for *Seq2* which are populated at equilibrium temperature are the same as the ones populated in the kinetic intermediate (shown in Fig. 6b).

Discussion

A detailed discussion of the role of folding intermediates was given in a recent paper by Fersht [10]. Arguments were presented that the evolutionary optimization of folding rate and stability was likely to press for elimination of folding intermediates. The lattice simulations presented in this work provide a complementary analysis and amplify these points.

The conclusion from our comparative analysis of folding of the two sequences is that the existence of intermediates does not facilitate folding but rather makes it slower and considerably decreases the thermodynamic stability of the native state. It is also clear that the very existence of intermediates is connected with the shortcomings of design ('evolutionary optimization' in the realm of the present model) which, for *Seq2*, was carried out under the restrictive condition of preserved amino acid composition. The character of unfolded and misfolded states is very sensitive to such 'averaged' properties of sequences as their amino acid composition [47]. Not surprisingly, when the improved design was applied, which optimized the relative energy, or the *Z*-score, as it was done for *Seq1*, it optimized both factors, i.e. decreased the energy of the native state E_{nat} and increased the average energy of the misfolded state E_{av}. The data on contact energetics for both sequences are given in Table 1.

Table 1

Energetic characteristics of *Seq1* and *Seq2*.

	E_{nat}	E_{av}	$E_{non\text{-}nat}$	σ	E_{rel}
Seq1	−0.364	0.041	0.106	0.288	−56.14
Seq2	−0.411	−0.015	0.05	0.35	−47.38

E_{nat} is the energy per one native contact. E_{av} is the average contact energy, calculated over all possible contacts. $E_{non\text{-}nat}$ is the average energy of non-native contact. σ is the standard variance of energies of all contacts. E_{rel} is defined in equation 2.

It is clear that the improved design, as well as making the energy of the native conformation somewhat higher than for *Seq2*, also 'designed out' the non-native conformations (making each non-native contact of *Seq1* ≈ 0.2 *T* — where *T*=0.26 is the temperature at which kinetic simulations were performed — higher in energy than of *Seq2*). This eliminated the intermediates, leaving only fast track and eliminating slower kinetic phases associated with strong non-native contacts in the intermediates.

What is the origin of fast and slow folding phases for *Seq2*? Our analysis clearly suggests that it is the heterogeneity of the intermediate that gives rise to the heterogeneity in folding pathways. Indeed, along with many strong native contacts, we found a few strong non-native contacts which are formed in the intermediate leading to its overall incorrect topology. It is these contacts that when formed give rise to a pronounced slow phase of folding. The fact that a small number of contacts are instrumental in generating the slow-folding phase suggests that it can be eliminated by a limited number of mutations. This was indeed observed in barnase, where the mutation I96→A made folding much closer to a two-state process [17]. The latter observation also explains why formation of burst compact intermediates so often results in kinetic traps. Indeed, formation of any stable contact is more likely when the chain is compact than when it is in the random coil state. Apparently, for any sequence there can be strong attractive interactions between amino acids that are not in contact in the native structure. Such contacts are more probable when the chain is compact.

Our findings for the lattice model can be summarized in the schematic landscape diagram for folding of *Seq1* and *Seq2* shown in Figure 13.

In this paper, we have presented a comprehensive analysis of the role of intermediates for folding of lattice model proteins. In previous instances, lattice models have been successful in identifying such key elements of the protein folding mechanism as a specific nucleus [10,12,53]. The important question remaining is to what extent are present results on the role of intermediates applicable to real proteins? In other words, how do the simulation results compare with experimental data?

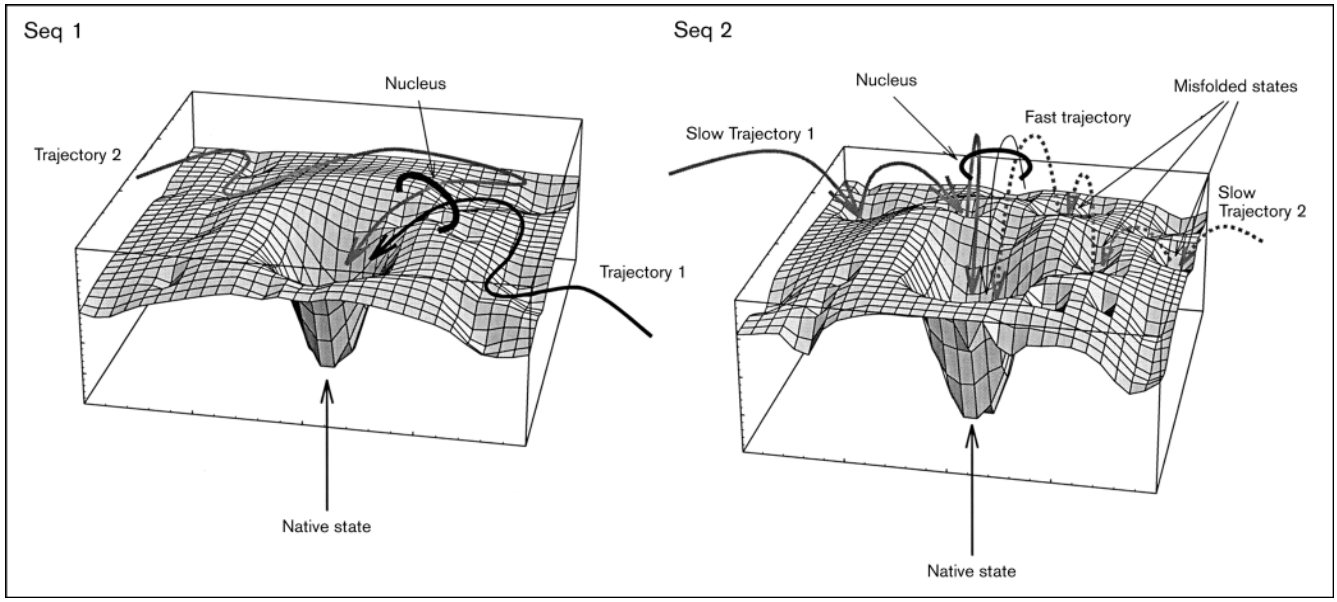
Experimentally, both scenarios described in this work (and summarized in Fig. 13) were observed for different proteins. Important examples of folding proceeding via intermediates are myoglobin [16], hen egg white lysozyme (HEWL) [18,58], barnase [17], cytochrome *c* at pH 7 [49], [59], and RNase A [50,51]. On the other hand, there are a growing number of examples where folding of a protein has been shown to follow a simple two-state scenario both in thermodynamics and in kinetics: chymotrypsin inhibitor 2 [19], ubiquitin at 8°C [22], cytochrome *c* at pH 5 [25], Ig-binding domain of staphylococcal protein G [20], SH3 [21], *E. coli* cold-shock protein [23], and acyl-coenzyme A binding protein [24].

Moreover, the same protein can exhibit, at different temperatures, either scenario [22,53], or a single mutation can cause a switch between folding mechanisms (e.g. mutation I96→A in barnase, see [17]). This is in accord with our findings for the lattice model in which we have shown that the mechanism is not determined by the native structure but by features of the sequence, as comparison between *Seq1* and *Seq2* suggests.

Another important experimental observation is that proteins which fold without intermediates are doing so faster than the ones which have intermediates on their pathway [19,20,22,23]. This trend is also in agreement with our results. However, one should bear in mind that in general, smaller proteins are known to fold without intermediates and therefore the factor of length should also be taken into account when comparing the folding rates of different proteins.

Kiefhaber showed in a recent work [58] that 14% of HEWL molecules follow the fast folding pathway (with time constant 50 ms) directly to the native state. This correspond to the fast stage of kinetics found in the earlier kinetic study of HEWL folding [18]. This is consistent with our analysis where partition into slow and fast folding trajectories in *Seq2* was found (see Fig. 13). An important result of the work of Kiefhaber [58] is that the folding rate on the fast pathway can be well approximated by the interpolation of the linear part of the dependence of log(folding rate) on final concentration of denaturant. This linear part corresponds to the regime when folding proceeds without intermediates, and its extrapolation estimates well the rate of the fast folding pathway. The dependence of log(folding rate) on final concentration of denaturant curves down at lower denaturant concentration due to a change of the folding mechanism to the one with

Figure 13



The schematic representation of the free energy landscape for *Seq1* and *Seq2*. Each trajectory passes through the saddle point which is the nucleation transition state.

folding intermediates [19,22]. This provides direct evidence that the existence of intermediates slows down the folding process.

The intermediates observed and characterized in this work (for *Seq2*) share a number of features with the molten globule state (reviewed by Ptitsyn in [60]). Indeed, a number of strong native-like contacts are present in experimentally observed molten globule intermediates for some proteins (e.g. see [16]), and our analysis shows the same for the lattice model. Further, we see that the intermediate (found in *Seq2*) has a number of pronounced non-native contacts. This is consistent with the recent observation that the subdomains most protected from hydrogen exchange are different in the molten globule state and in the native state [61].

Our simulations clearly suggest that (for *Seq2*) kinetic and equilibrium intermediates, when present, have similar structural features. This points to another consistency between intermediates studied in this work and molten globule like folding intermediates found in experiments. Jennings and Wright [16] studied the structure of the 5 ms folding intermediate of myoglobin and found that it is similar to the structure of equilibrium molten globule at low pH [62]. Similar observations were made for the acid-induced form of cytochrome *c* [49,63].

The results of this analysis have a number of implications for future experimental and theoretical studies. The 'molten globule' and framework models of folding convinced many experimentalists that the most relevant events take place at fast and ultrafast stages of folding. A similar point of view was held by some theoreticians, but for a different reason: that at slow stages, 'adventitious' traps are mainly found and therefore their study is not so relevant. Our analysis (see also a very detailed discussion by Fersht [10]) suggests that although slower stages of folding can indeed be due to trapping mechanisms, the cooperative character of folding makes ultrafast events (much faster than the time constant of the fastest stage of folding to the native state [58]) not really relevant for folding either. It is most likely that at these time scales all that are taking place are stochastic fluctuations in the unfolded state, while protein progresses to the native structure only after the nucleation free energy barrier is overcome. This is especially clear for proteins that have two-state folding kinetics, as does *Seq1* in this study. Indeed at any time interval the only species that can be found are unfolded and folded ones, and all that is changing with time is their proportion (see Fig. 10).

In this work, we studied relatively short sequences. The transition state for folding of such short lattice model chains as well as small proteins involves formation of a specific and unique nucleus, i.e. the set of contacts whose formation is necessary and sufficient for subsequent fast folding to the native state [12,53]. It was pointed out in our earlier publication [12] that longer sequences may have multiple nuclei and that their intermediates may correspond to the cases in which one folding nucleus is formed while the other(s) are not. Such proteins will fold via a multidomain mechanism. This was indeed found in simulations of longer chain folding [54] and it also follows from experimental data on HEWL [64], barnase [53] and staphylococcal nuclease [65]. More complicated folding scenarios taking place in longer chains will be the subject of future study.

Note added in proof

After this work had been completed, we became aware of an interesting article by Radford and Dobson [66] in which parallel pathways and partitioning into slow and fast folding trajectories are discussed from an experimental perspective. The experimental results reported are in good agreement with the results of the simulations we present here.

Acknowledgement

This work was supported by Packard Foundation and grant GM-52126 from NIH.

References

1. Levinthal, C. (1968). Are there pathways for protein folding? *J. Chim. Phys.* **65**, 44.
2. Ptitsyn, O.B. (1973). Stagewise mechanism of protein molecule self-organization. *Vestn. Akad. Nauk. SSSR* **5**, 57–62.
3. Kim, P. & Baldwin, R. (1982). Specific intermediates in the folding reaction of small proteins and the mechanism of protein folding. *Annu. Rev. Biochem.* **51**, 459.
4. Karplus, M. & Weaver, D. (1976). Protein-folding dynamics. *Nature* **160**, 404–406.
5. Landau, L.D. & Lifshitz, E.M. (1980). *Statistical Physics.* Pergamon, London.
6. Lifshitz, E.M. & Pitaevski, L.P. (1981). *Physical Kinetics.* Pergamon, London.
7. Karplus, M. & Shakhnovich, E.I. (1992). Protein folding: theoretical studies of thermodynamics and dynamics. In *Protein Folding*, pp. 127–195. W.H. Freeman and Company, New York.
8. De Gennes, P.G. (1985). Kinetics of collapse of flexible coil. *J. Physique Lett.* **46**, L639–641.
9. Grosberg, A.Y., Nechaev, S.K. & Shakhnovich, E.I. (1988). The role of topological constraints in the kinetics of collapse of macromolecules. *J. Physique (France)* **49**, 2095–2100.
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. Matouschek, A., Kellis J., Jr, Serrano, L., Bycroft, M. & Fersht, A.R. (1990). Transient folding intermediates characterized by protein engineering. *Nature* **346**, 440–445.
12. Abkevich, V.I, Gutin, A.M. & Shakhnovich, E.I. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* **33**, 10026–10036.
13. Privalov, P.L. & Khechinashvili, N.N. (1974). A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. *J. Mol. Biol.* **86**, 665–679.
14. Privalov, P.L. (1979). Stability of proteins. Small single-domain proteins. *Adv. Protein Chem.* **33**, 167–241.
15. Lifshitz, I.M., Grosberg, A.Y. & Khohlov, A.R. (1978). Some problems of statistical physics of polymers with volume interactions. *Rev. Mod. Phys.* **50**, 683–713.
16. Jennings, P. & Wright, P. (1993). A molten globule intermediate formed early on the kinetic folding pathway of myoglobin. *Science* **262**, 892–896.

17. Matouschek, A., Serrano, L. & Fersht, A.R. (1992). The folding of an enzyme. iv the structure of an intermediate in the refolding of barnase analyzed by protein engineering procedure. *J. Mol. Biol.* **224**, 819–835.
18. Radford, S., Dobson, C. & Evans, P. (1992). The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature* **358**, 302–307.
19. Jackson, S.E. & Fersht, A.R. (1991). Folding of chimotrypsin inhibitor 2. 1. Evidence for a two-state transition. *Biochemistry* **30**, 10428–10435.
20. Alexander, P., Orban, J. & Bryan, P. (1992). Kinetic analysis of folding and unfolding of 56 amino acid IGG-binding domain of streptococcal protein G. *Biochemistry* **31**, 7243–7248.
21. Viguera, A.R., Martinez, J.C., Filimonov, V.V., Mateo, P.L. & Serrano, L. (1994). Thermodynamic and kinetic analysis of the SH3 domain of spectrin shows a two-state folding transition. *Biochemistry* **33**, 2142–2150.
22. Khorasanizadeh, S., Peters, I.D. & Roder, H. (1993). Folding and stability of a tryptophan-containing mutant of ubiquitin. *Biochemistry* **32**, 7054–7063.
23. Schindler, T., Herrler, M., Marahiel, M. & Schmid, M. (1995). Extremely rapid protein folding in the absence of intermediates. *Nature Struct. Biol.* **2**, 663–673.
24. Kragelund, B.B., Robinson, C.V., Knudsen, J. & Dobson, C.M. (1995). Folding of a four-helix bundle: studies of acetyl-coenzyme A binding protein. *Biochemistry* **34**, 7117–7124.
25. Sosnick, T.R., Mayne, L., Hiller, R. & Englander, S. (1994). The barriers in protein folding. *Nature Struct. Biol.* **1**, 149–156.
26. Bai, Y., Sosnick, T.R., Mayne, T. & Englander, S.W. (1995). Protein folding intermediates: native state hydrogen exchange. *Science* **269**, 192–197.
27. Khorasanizadeh, S., Peters, I.D., Butt, T.R. & Roder, H. (1996). Evidence for a general three-state model of protein folding from kinetic analysis of ubiquitin variants with altered core residues. *Nature Struct. Biol.* in press.
28. Gutin, A.M., Abkevich, V.I. & Shakhnovich, E.I (1995). Is burst hydrophobic collapse necessary for rapid folding? *Biochemistry* **34**, 3066–3076.
29. Skolnick, J. & Kolinski, A. (1991). Dynamic monte carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *J. Mol. Biol.* **221**, 499–531.
30. Shakhnovich, E.I., Farztdinov, G.M., Gutin, A.M. & Karplus, M. (1991). Protein folding bottlenecks: a lattice monte-carlo simulation. *Phys. Rev. Lett.* **67**, 1665–1667.
31. Miller, R., Danko, C., Fasolka, M.J., Balazs, A.C., Chan, H.S. & Dill, K.A. (1992). Folding kinetics of proteins and copolymers. *J. Chem. Phys.* **96**, 768–780.
32. Camacho, C. & Thirumalai, D. (1993). Kinetics and thermodynamics of folding in model proteins. *Proc. Natl. Acad. Sci. USA* **90**, 6369–6372.
33. Kolinski, A., Godzik, A. & Skolnick, J. (1993). The general method for the prediction of the three-dimensional structure and folding pathway of globular proteins: application to designed helical proteins. *J. Chem. Phys.* **98**, 7420–7433.
34. Sali, A., Shakhnovich, E.I. & Karplus, M. (1994). Kinetics of protein folding: a lattice model study for the requirements for folding to the native state. *J. Mol. Biol.* **235**,1614–1636.
35. Shakhnovich, E.I. (1994). Proteins with selected sequences fold to their unique native conformation. *Phys. Rev. Lett.* **72**, 3907–3910.
36. Socci, N. & Onuchic, J.N. (1994). Folding kinetics of protein like heteropolymers. *J. Chem. Phys.* **101**, 1519–1528.
37. Chan, H.S. & Dill, K.A. (1994). Transition states and folding dynamics of proteins and heteropolymers. *J. Chem. Phys.* **100**, 9238–9257.
38. Myazawa, S. & Jernigan, R. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534–552.
39. Goldstein, R., Luthey-Schulten, Z.A. & Wolynes, P. (1992). Optimal protein-folding codes from spin-glass theory. *Proc. Natl. Acad. Sci. USA* **89**, 4918–4922.
40. Abkevich, V.I., Gutin, A.M. & Shakhnovich, E.I. (1994). Free energy landscape for protein folding kinetics. Intermediates, traps and multiple pathways in theory and lattice model simulations. *J. Chem. Phys.* **101**, 6052–6062.
41. Bryngelson, J., Onuchic, J.N., Socci, N.D. & Wolynes, P. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **21**, 167–195.
42. Shakhnovich, E.I. & Gutin, A.M. (1993). Engineering of stable and fast-folding sequences of model proteins. *Proc.Natl. Acad. Sci. USA* **90**, 7195–7199.
43. Shakhnovich, E.I & Gutin, A.M. (1993). A novel approach to design of stable proteins. *Protein Eng.* **6**, 793–800.
44. Abkevich, V.I., Gutin, A.M & Shakhnovich, E.I. (1995). Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *J. Mol. Biol.* **252**, 460–471.
45. Creighton, T. (1992). *Proteins. Structure and Molecular Properties.* W.H. Freeman & Co, New York.
46. Bowie, J.U., Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164–169.
47. Gutin, A.M., Abkevich, V.I. & Shakhnovich, E.I. (1995). Evolution-like selection of fast-folding model proteins. *Proc Natl. Acad. Sci. USA* **92**, 1282–1286.
48. Rey, J. & Skolnick, J. (1991). Comparison of lattice monte-carlo and brownian dynamics folding pathways of α-helical hairpins. *Chem. Phys.* **158**, 199–212.
49. Elove, G., Roder, H. & Englander, S. (1988). Structural characterization of folding intermediates in cytochrome c by H-exchange labeling and proton NMR. *Nature* **335**, 700–704.
50. Udgaonkar, J. & Baldwin, R. (1988). NMR evidence for an early framework intermediate on the folding pathway of the ribonuclease A. *Nature* **335**, 694–700.
51. Baldwin, R. (1993). Pulsed h/d exchange studies of folding intermediates. *Curr. Opin. Struct. Biol.* **3**, 84–91.
52. Guo, Z. & Thirumalai, D. (1995). Nucleation mechanism for protein folding and theoretical predictions for hydrogen-exchange labelling experiments. *Biopolymers* **35**, 137–139.
53. Itzhaki, L. Otzen. D. & 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.
54. Abkevich, V.I., Gutin, A.M. & Shakhnovich, E.I. (1995). Domains in folding of model proteins. *Protein Sci.* **4**, 1167–1177.
55. Ferrenberg, A.M. & Swendsen, R.H. (1989). Optimized Monte Carlo data analysis. *Phys. Rev. Lett.* **63**, 1195–1197.
56. Sali, A., Shakhnovich, E.I. & Karplus, M. (1994). How does a protein fold? *Nature* **369**, 248–251.
57. Socci, N. & Onuchic, J. (1995). Kinetics and thermodynamic analysis of proteinlike heteropolymer: Monte Carlo histogram technique. *J. Chem. Phys.* **103**, 4732–4744.
58. Kiefhaber, T. (1995). Kinetic traps in lysozyme folding. *Proc. Natl. Acad. Sci USA* **92**, 9029–9033.
59. Elove, G., Charlotte, A., Roder, H. & Goldberg, M. (1992). Early steps in cytochrome c folding probed by time-resolved circular dichroism and fluorescence spectroscopy. *Biochemistry* **31**, 6876–6883.
60. Ptitsyn, O.B. (1992). The molten globule state. In: *Protein Folding,* pp. 243–300. W.H. Freeman and Company, New York.
61. Schulman, B.A., Redfeld, C.A., Peng, Z., Dobson, C.M. & Kim, P.S. (1995). Different subdomains are most protected from hydrogen exchange in the molten globule and native states of human α-lactalbumin. *J. Mol. Biol.* **253**, 651–657.
62. Hughson, F., Wright, P. & Baldwin, R. (1990). Structural characterization of the partly folded apomyoglobin intermediate. *Science* **249**, 1544–1548.
63. Jeng, M.F., Englander, S.W., Elove, G., Wand, A. & Roder, H. (1990). Structural description of acid-denatured cytochrome c by hydrogen exchange and 2D NMR. *Biochemistry* **29**, 10433–10437.
64. Miranker, A., Radford, S., Karplus, M. & Dobson, C. (1991). Demonstration by NMR of folding domains in lysozyme. *Nature* **349**, 633–636.
65. Carra, J.H., Anderson, E.A. & Privalov, P.L. (1994). Three-state thermodynamic analysis of the denaturation of staphylococcal nuclease mutants. *Biochemistry* **33**, 10842–10850.
66. Radford, S.E. & Dobson, C.M. (1995). Insights into protein folding using physical techniques: studies of lysozyme and alpha-lactalbumin. *Philos. Trans. R. Soc. Lond. B* **348,** 17–25.

Because *Folding & Design* operates a 'Continuous Publication System' for Research Papers, this paper will have been published via the internet before being printed. For information on how to gain access via the internet, see the explanation on the contents page.