doi:10.1006/jmbi.2001.4586 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2001) **308**, 79–95

JMB

The Folding Thermodynamics and Kinetics of Crambin Using an All-atom Monte Carlo Simulation

Jun Shimada, Edo L. Kussell and Eugene I. Shakhnovich*

Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street Cambridge, MA 02138, USA

**Corresponding author*

We present a novel Monte Carlo simulation of protein folding, in which all heavy atoms are represented as interacting hard spheres. This model includes all degrees of freedom relevant to folding, all side-chain and backbone torsions, and uses a Gō potential. In this study, we focus on the 46 residue α/β protein crambin and two of its structural components, the helix and helix hairpin. For a wide range of temperatures, we recorded multiple folding events of these three structures from random coils to native conformations that differ by less than 1 Å C^{α} dRMS from their crystal structure coordinates. The thermodynamics and kinetic mechanism of the helix-coil transition obtained from our simulation shows excellent agreement with currently available experimental and molecular dynamics data. Based on insights obtained from folding its smaller structural components, a possible folding mechanism for crambin is proposed. We observed that the folding occurs *via* a cooperative, first order-like process, and that many folding pathways to the native state exist. One particular sequence of events constitutes a "fast-folding" pathway where kinetic traps are avoided. At very low temperatures, a kinetic trap arising from the incorrect packing of side-chains was observed. These results demonstrate that folding to the native state can be observed in a reasonable amount of time on desktop computers even when an all-atom representation is used, provided the energetics sufficiently stabilize the native state.

Keywords: protein folding; Monte Carlo; all-atom simulations; nucleation-condensation; crambin

Introduction

Previous lattice and off-lattice folding simulations featured coarse-grained protein representations where amino acids were often modeled as spheres (Sali *et al.*, 1994; Shakhnovich, 1997; Pande *et al.*, 1998; Dill *et al.*, 1995; Bryngelson *et al.*, 1995; Klimov & Thirumalai, 1996; Honeycutt & Thirumalai, 1992; Berriz *et al.*, 1996; Zhou & Karplus, 1999; Clementi *et al.*, 2000). While such simplified representations have their advantages, such as the reduction in the number of degrees of freedom, there are several shortcomings. First, one could argue that such models do not capture the full complexity of conformational space, thereby not truly addressing the Levinthal problem (Creighton, 1992). Second, coarse-grained models may lack realistic features of secondary structure elements. Finally, such models do not address the packing of side-chains in the protein interior, which is often viewed as an important aspect of the folding process, given the diversity of side-chain shapes and their dense packing in the native state (Richards & Lim, 1994).

Molecular dynamics approaches have attempted to bridge the gap between simulation and reality by using atomic representations of proteins with explicit solvent molecules. Unfortunately, the computational times for such minimally coarse-grained simulations are still prohibitively high as evidenced by a recent work that yielded only a partial folding trajectory (Duan & Kollman, 1998) even with the aid of a massively parallel supercomputer. An ensemble of complete folding trajectories is required in order to gain meaningful physical insights, especially as our theoretical understanding of protein folding has become increasingly grounded on statistical mechanical principles (Shakhnovich, 1997; Onuchic *et al.*, 1997; Pande

Abbreviations used: MC, Monte Carlo; FPT, first passage time.

E-mail address of the corresponding author: eugene@belok.harvard.edu *et al.*, 2000). Other molecular dynamics research efforts have tried to address this issue by obtaining multiple unfolding runs (Li & Daggett, 1996; Lazaridis & Karplus, 1997) at extremely high temperatures. Unfortunately, it is unclear whether the observed unfolding pathways are indicative of folding pathways under normal conditions (Finkelstein, 1997; Dinner & Karplus, 1999).

We present a Monte Carlo (MC) simulation (Binder & Heerman, 1992) which combines an all-atom description of the protein with coarse-grained motions and energetics. In this simulation, (1) all heavy atoms in the protein are represented as impenetrable spheres, (2) all backbone and side-chain torsions, which account for all degrees of freedom relevant to folding, are allowed to move, and (3) a square well, Gō potential (McQuarrie, 1976; Gō & Abe, 1981) is used for the interaction energy. There are several advantages to this method. First, a statistically significant number of complete folding trajectories can be collected using conventional computational resources. Second, the atomic-level resolution of the simulation yields detailed descriptions of the folding process of actual protein structures including side-chain packing. Finally, this coarse-grained approach allows for a systematic investigation of the physical principles dictating protein folding. If one begins with a detailed energy function as in molecular dynamics, it may be difficult to deconvolute exactly which energy terms are essential for folding. On the other hand, as demonstrated in theoretical investigations (Grosberg & Khokhlov, 1994), by thoroughly investigating the successes and limitations of coarse-grained models, it is possible to test which features of a model are necessary and/or sufficient for describing the complex physics of heteropolymers.

Using our simulation, we have repeatedly folded the 46-residue α-β protein crambin (Figure 1) to within 1 Å backbone dRMS). Furthermore, the thermodynamic and kinetic folding properties obtained from our simulation are consistent with experimental studies of other single domain proteins (Grantcharova & Baker, 1997; Jackson & Fersht, 1991), for which the folding transition generally exhibits two-state behavior with no accumulating intermediates between the denatured and native states (Jackson, 1998).

One of the more successful methods reported in the literature generated folded structures for crambin to ≈3 Å C^{α} dRMS (Kolinski & Skolnick, 1994) using a sequence-based potential. These final structures were obtained from a two-step heuristic approach: simplified structures obtained from folding runs on a coarser lattice were then refined on a finer lattice. In contrast, our method uses a potential based on knowledge of the native structure, but generates the entire folding transition without altering key conditions (such as the move set, temperature, or protein representation) once the simulation is initiated. As such, it produces an ensemble of trajectories which can yield thermodynamic and kinetic information. Although the potential we employ cannot be used to fold arbitrary protein sequences, we demonstrate that the full conformational search problem can be solved using standard computational resources.

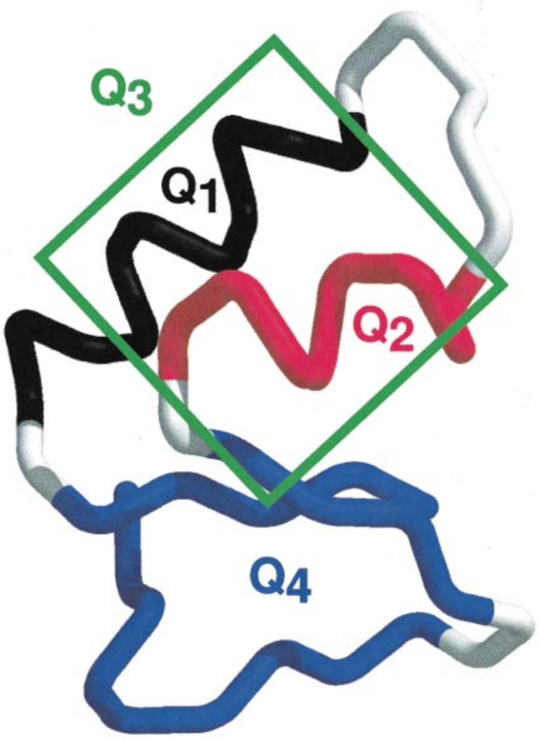

Figure 1. The 46-residue protein crambin. The important secondary/tertiary structural elements are indicated by different colors: black, helix 1 (residues 6-18, sequence: SIVARSNFNVCRL); red, helix 2 (23-30, EAICATYT); green, inter-helix contacts; blue, β-sheet (1-5, 32-35, 38-46). As shown by the matching colors, the Q_i parameters are defined as the fraction of native contacts in specific structural elements: Q_1, helix 1; Q_2, helix 2; Q_3, inter-helix contacts; Q_4, β-sheet.

Theory and Motivation

Analytical studies (Bryngelson & Wolynes, 1987; Ramanathan & Shakhnovich, 1994; Shakhnovich, 1997; Pande *et al.*, 2000) and lattice simulations (Shakhnovich & Gutin, 1993) determined that only certain sequences can fold in a biologically relevant time scale. Such fast-folding sequences featured the native conformation as the pronounced energy minimum. For example, when evolution-like pressures were applied to lattice protein sequences to preferentially select mutations that improved folding speed (analogous to real-life evolutionary pressures to select peptide sequences that survive proteolysis), the resulting sequences featured a large energy gap between their native and competing non-native states (Gutin *et al.*, 1995). Conversely, in order to successfully simulate the folding of biologically occurring protein sequences, it is necessary to use a model that places the native state at the bottom of the energy spectrum.

As noted in the Introduction, our folding model combines a near-perfect geometrical representation of the peptide chain with a potential energy function that delivers the required ''energy gap'' by explicitly assigning the native conformation as the ground state (see Methods). In general, a Gō potential is often viewed as providing an idealized energy landscape by strongly biasing folding: first, there is a strong correlation between energy and structural distance (e.g. C^{α} RMS) from the native state; and second, the potential energy surface is seen as very smooth and downhill when going from the unfolded to the native state.

Despite this criticism, important aspects about folding kinetics can be learned as demonstrated by several recent studies (Pande & Rokhsar, 1999; Zhou & Karplus, 1999; Clementi *et al.*, 2000). One particularly interesting aspect is the role topology may play during the folding process (Baker, 2000). Frustration due to topology may be problematic for long enough chains, particularly when nonlocal geometrical constraints (e.g. β-sheets) must be met while satisfying local ones (e.g. α-helices). For simulations using Gō energetics, this feature can be captured if there are excluded volume interactions (thereby prohibiting chain crossings) and realistic backbone conformations. Indeed, it is demonstrated below that after a relatively downhill compactization process, the folding process proceeds on a rugged free energy landscape. The presence of severe kinetic traps due to topological frustration results in preferred folding pathways and thus makes the ensemble kinetics extremely complex.

The move set

To a good approximation, the conformational motions of a solvated polymer at room temperature occur *via* the torsional degrees of freedom (Grosberg & Khokhlov, 1994). Our main concern when choosing the move set was to balance realism and efficiency. On the one hand, a complicated move set (e.g. a loop move of *n* residues while keeping the ends fixed) would result in computational inefficiencies. On the other, using only single torsional rotations as the move set is unrealistic, since it may frequently result in large conformational changes of the polymer chain. Such large moves are generally prohibited in solvent, because of the drag experienced by the moving chain, and in the compact state, because of excluded volume restrictions.

We therefore opted for a simple torsional move set with two important features. First, all of our moves consist of either two, four or six concerted rotations. This allows for a variety of ''loop''-like moves, which are particularly important when in the compact state. These rotations were required to be within six residues of each other, in order to ensure that the majority of moves resulted in local conformational changes. Second, the step sizes for these concerted rotations were drawn from a Gaussian distribution of relatively small width. More specificially, the probability for selecting a backbone step size of θ was given by:

$$p(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\theta^2}{2\sigma^2}\right)$$

where $\sigma = 2^{\circ}$. The vast majority of step sizes are small enough that the conformational changes induced by our moves are local. Very rarely, when a fortuitous combination of torsions and rotations are chosen and the chain is in an extended state, a global conformational change is allowed. During all of our simulations, such global conformational changes were not observed in the compact state: large conformational changes were always accompanied by partial unfolding events. In this manner, by allowing primarily global moves in the extended state and loop moves with small step sizes in the compact state, our simulation allows for efficient sampling of conformational space. In addition, the acceptance ratio drops dramatically from >90 % to >10 % upon compactization from the coiled state, which agrees with our physical intution that polymer dynamics should dramatically slow down after collapse.

Can our Monte Carlo simulation yield information about folding kinetics?

MC simulations have been used in a wide range of problems in statistical physics to obtain equilibrium information (Binder & Heerman, 1992). As long as a MC simulation's move set satisfies the criterion of detailed balance, convergence to a canonical distribution is guaranteed. A much more controversial aspect of MC simulations has been its use in determining the relaxation kinetics to an equilibrium distribution. Because there is no explicit correspondence with real time in a MC simulation, one might raise the objection that all kinetic interpretations of MC trajectories are dependent on the particular move set chosen.

For simple MC lattice folding simulations utilizing the standard Verdier move set (consisting of the crankshaft, corner and tail flips) (Verdier, 1973), it was demonstrated that general conclusions on folding kinetics, such as the importance of specific nucleation for fast-folding, could be drawn if the analysis was performed on an ensemble of folding runs (Abkevich *et al.*, 1994a,b). This approach was supported by an eariler study, which analytically showed that the coil-globule relaxation time, as computed by the use of the Verdier and other local move sets, matched theoretical predictions (Hilhorst & Deutch, 1975). Unfortunately, such rigorous analyses rapidly become difficult as the complexity of the system being simulated increases.

To justify our use of MC to interpret folding kinetics, we provide the following heuristic argument. Let $\vec{W}(s) = \{w_i(s)\}$ represent the probability distribution of states at step s of a MC simulation. We can then represent the progression in our MC simulation as the Markov process:

$$\vec{W}(s+1) = P \times \vec{W}(s)$$

where P is the transition matrix, whose $i-j$th element is given by the probability to go from state i to j. In a Metropolis MC simulation, the individual elements of this transition matrix obey detailed balance, which means it is the equilibrium solution to the series of first-order kinetic equations (also known as the master equations):

$$\frac{dw_i}{dt} = \sum_j -P_{ij}w_i + P_{ji}w_j = 0$$

Now consider a move set that is sufficiently local and unbiased towards the native state. During the relaxation to equilibrium, the ensemble population will be highest where there are free energy minima. The evolution of the MC simulation thus leads to a gradual shifting of the ensemble population from local minima to the global minimum, the native state. If kinetic events are separated by a large enough number of local moves and they are observed in an ensemble of relaxation trajectories, they represent significant state population shifts and reflect properties of the free energy landscape (e.g. the depth of local minima and barrier heights). In this manner, examining an ensemble of trajectories highlights the major kinetic events during folding by averaging out short MC time events. Importantly, this has the effect of minimizing differences arising from the selection of a particular move set.

The kinetic picture arising from a MC simulation is thus coarse-grained. If a common free energy landscape is used, both MC and other types of simulations should agree on the sequence of major folding events, if they are sufficiently separated in time. This was demonstrated by Rey & Skolnick (1991), who concluded that a MC simulation yielded the same folding pathways for an α-helical hairpin as a Brownian dynamics simulation, for which there was real time evolution. We emphasize that MC simulations are not immediately justified in making quantitative predictions or microscopic analyses of folding kinetics. In order to reliably predict folding rates, every MC simulation needs to be extensively calibrated to existing experimental data. It is likely that such calibration will rule out several move sets.

Finally, perhaps the most important requirement of achieving quantitative agreement with experimental kinetic data is that the free energy landscape must be sufficiently realistic. Even molecular and Langevin dynamics simulations, where a well-defined time step is used to evolve the simulation, will be unable to make quantitative, and perhaps even qualitative, analyses of folding kinetics if this requirement is not met. Because there have been no reports of a successful *ab initio* folding potential, we decided that the Gō potential, combined with an all-atom representation of the protein, was the best choice for this simulation.

Before proceeding to analyze the folding kinetics of crambin, we decided to more accurately assess the validity of our move set by examining the thermodynamics and kinetics of the helix-coil transition, for which there are experimental and molecular dynamics data.

Results

The formation of the $(i,i+4)$ hydrogen bonding scheme in helical structures results in a high native contact density. As a result, a non-discriminate use of Gō potentials will lead to unusually strong energetic biases towards helices. Since we were interested in probing folding mechanisms under physiological conditions, we calibrated the stability of the two crambin helices with the predictions of AGADIR (Muñoz & Serrano, 1994) (at pH = 7 and $T = 298$ K). As for most peptide sequences (Scholtz & Baldwin, 1992; Muñoz & Serrano, 1995), both helices were reported to have zero helical propensities at temperatures and pH relevant for our folding simulations. Elimination of backbone contacts was the simplest way to dramatically reduce helix stability without introducing biases due to sequence identity. All subsequent data were collected while using this slightly modified Gō potential.

Helix-coil transition

Both helix 1 and 2 (see Figure 1 for sequences) exhibited an abrupt transition to the coil state, with the transition temperatures occurring at $T_f^{\mathrm{helix1}} \approx 1.2$ and $T_f^{\mathrm{helix2}} \approx 0.9$, respectively (Figure 3(a) and (b)). Because we observed that the thermodynamic and kinetic data are qualitatively the same for the two helices aside from minor details (such as this shift in the transition temperature), we focused our analysis on helix 1.

The energy histograms show a sudden shift in helix and coil state populations as T_f^{helix1} is passed, with a broadening of the distribution at T_f^{helix1} (Figure 2). The free energy curves clearly demonstrate the emergence of a local, second minimum centered at 3.5 Å backbone dRMS for $1.1 < T < T_f^{\mathrm{helix1}}$ (Figure 3(c)). At T_f^{helix1}, this second minimum becomes equal in free energy to the helix state minimum, resulting in an equilibrium of both helix and coiled states. Finally, the heat capacities for both helices are sharply peaked around T_f (Figure 3(d)). Based on the width of the heat capacity peak at T_f, it appears that helix 2 shows a weaker transition, which is expected for a shorter helix. These thermodynamic observations fully agree with Zimm-Bragg theory (Zimm & Bragg, 1959), simulations (Daggett & Levitt, 1992; Daura *et al*., 1999; Hummer *et al*., 2000), and experiment (Thompson *et al*., 1997).

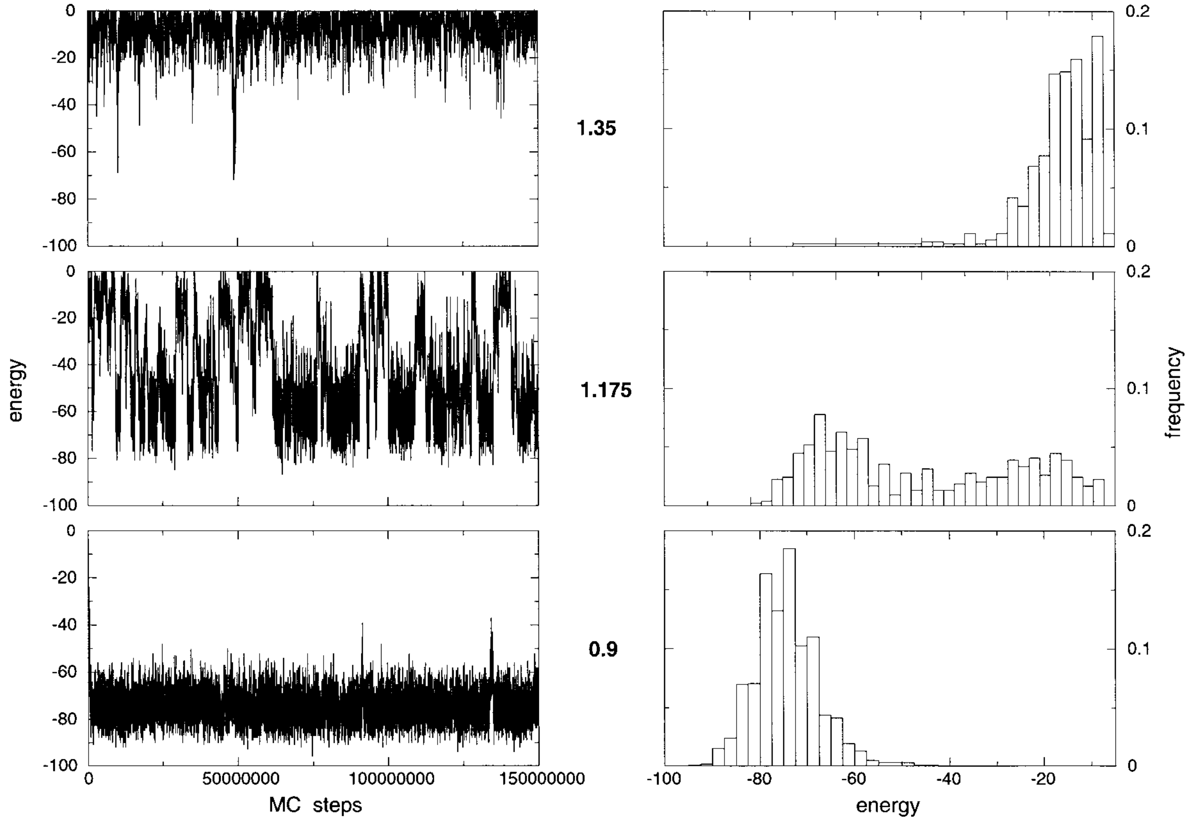

Figure 2. Helix-coil transition trajectories (left panels) and energy histograms (right panels) for helix 1 at various temperatures: above ($T = 1.35$; top panels), near ($T = 1.175$; middle panels), and below ($T = 0.9$; bottom panels) transition temperature. In this and all subsequent Figures, T is given in reduced energy units.

The median first passage time (FPT) for helix 1 rose rapidly as T was increased for $T > 1.1$ (Figure 4(a)). Given that the Gō energy more or less increases with increasing dRMS, it is clear that the free energy barrier to helix formation (Figure 3(c)) is purely entropic. Interestingly, for $T < 1.1$, the helix 1 formation rate becomes weakly dependent on temperature, suggesting that there are no major free energy barriers. At all temperatures, the distribution of FPTs exhibited long tails, but more data was needed to accurately determine whether the distribution was non-exponential. As observed in other simulations, the formation rate appears to be dominated by a diffusive search for a nucleation event (Hummer *et al.*, 2000, 2001). Once the nucleus is formed, the formation of the rest of the helix is rapid (see the trajectory at $T \approx T_f$ in Figure 2).

Examination of the ensemble helix 1 formation kinetics shows that Ala9 is the dominant nucleation site at both low and high temperatures (Figure 4(b)-(e)). This is shown by the rise in the fraction of native contacts (Q) made by the amide nitrogen of Ala9. It is necessarily accompanied by rise in the carbonyl oxygen Q of Ser6, as Ala9 and Ser6 make key side-chain/side-chain contacts to initiate the formation of the first turn. A logical explanation for this dominant nucleation event is the absence of side-chain entropy at Ala9 in our model, thus lowering the free energy cost of constraining this residue in a helical conformation. Upon nucleation, at higher temperatures, the propagation towards the C terminus (Figure 4(c) and (e)) appears to be stalled at Arg10 while at lower temperatures, this behavior is not observed (Figure 4(b) and (d)). This may be explained by the larger entropic cost of constraining the large side-chain of Arg10 at higher temperatures. A similar observation was made by Pande *et al.* (personal communication, 2000), whose simulation showed that helix propagation was slowed by the presence of arginine residues in a poly-alanine helix. The two exponential relaxation processes detected in the A_8-R-A_4 peptide by Thompson *et al.* (1997) may be explained by similar phenomena: the fast quenching of the N terminus fluorescent label is due to the rapid nucleation/propagation of the alanine-rich regions (from both N and C termini) and the slow phase due to the slow incorporation of the arginine residue into the growing helix.

At lower temperatures, Ser11 appears as a competing nucleation site. Like Ala9, Ser11 is a low entropy residue. However, in order for Ser11 to make helical contacts with neighboring residues, it requires constraining at least one of three large side-chains (Arg10, Asn12, Phe13), which may be

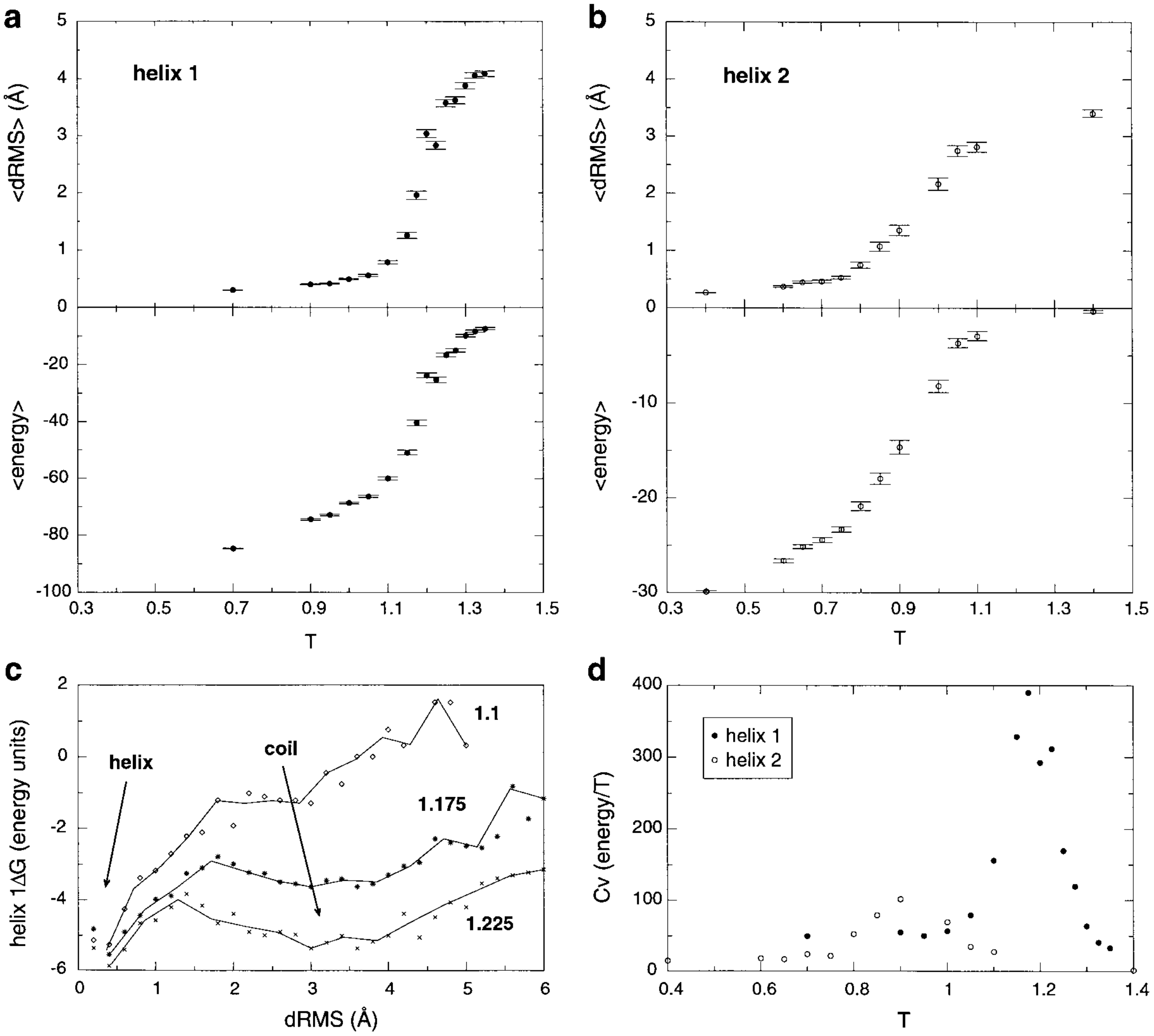

Figure 3. Summary of helix-coil transition thermodynamics. (a), (b) Transition curves for average E and backbone dRMS for helices 1 and 2, respectively. Note the difference in scales along the y-axis. (c) Free energy of helix 1 formation at various temperatures (see Methods). The backbone dRMS is chosen as the order parameter. Near the transition temperature ($T = 1.175$-1.225), two distinct free energy minima appear (labeled helix and coil). (d) Heat capacity (C_v) curves for helices 1 and 2. C_v was computed *via* the relation, $C_v = (\langle E^2\rangle - \langle E\rangle^2)/T^2$, where $\langle\rangle$ refers to ensemble averages (McQuarrie, 1976).

unfavorable for entropic or steric reasons. In contrast, forming the Ser6-Ala9 interaction requires constraining smaller residues. For this reason, it seems to be more favorable to nucleate the helix at Ala9. In support of this idea, we note that Ser11 disappears as a nucleation site at higher temperatures.

A third nucleation site occurs at Leu18. This appears to be driven primarily by energy, as many native side-chain/side-chain contacts are made with Val15. However, as the effect of side-chain entropy is increased as temperature is raised, we see that the Val15-Leu18 interaction forms slower relative to the Ser6-Ala9 interaction than at lower temperatures. Interestingly, raising the temperature seems to have a larger effect in reducing the capacity of Ser11 to nucleate the helix compared to Leu18.

Although the propagation asymmetry predicted by Young & Brooks (1996) was not observed at all temperatures, their general conclusion seems to be correct: preferences in helix propagation and nucleation are determined by the specific atom-atom interactions.

Folding simulations of crambin

Given that our move set yielded reasonable kinetic and thermodynamic helix-coil transition data, we proceeded to fold crambin. Starting from random coils, we considered the protein folded if the following three criteria were satisfied for at least 5×10^6 steps: (1) the fraction of native contacts (Q) exceeded 0.7; (2) the backbone dRMS was less than 1.25 Å; (3) the fraction of native contacts in the four secondary/tertiary structural features (Q_i,

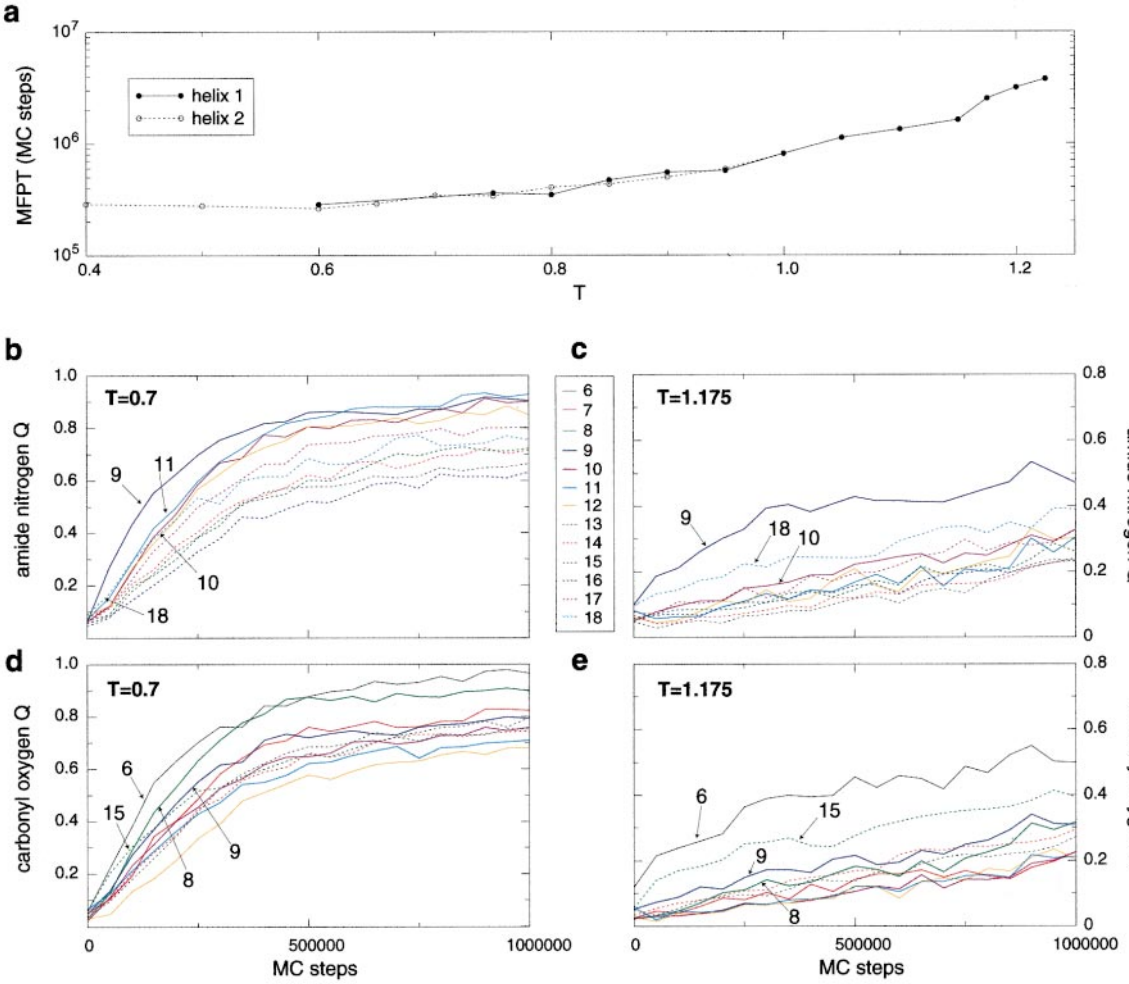

Figure 4. Summary of helix-coil transition kinetics. (a) Median first passage time (FPT) for helices 1 and 2. (b)-(e) Ensemble averaged kinetics of the early events during helix 1 formation at $T = 0.7$ and $T = 1.175$. This represents the average fraction (Q) of backbone amide ((b) and (c)) and carbonyl ((d) and (e)) native contacts observed in 100 runs as a function of MC steps. An increase in the amide Q reflects helix formation towards the N terminus, while an increase in the carbonyl Q reflects helix formation towards the C terminus. Residues that show significant rises in Q are labeled. Note the difference in scales along the y-axes.

$1 \leqslant i \leqslant 4$) exceeded 0.5 (see Figure 1). Criterion (1), which is similar to the one used to signify a folding event in lattice simulations (Abkevich *et al.*, 1994a), could not by itself distinguish conformations that one might intuitively consider folded (i.e. properly formed secondary structure and low overall dRMS) from obvious misfolds. Although our definition for folding did not measure when equilibrium was attained, it was useful for identifying when the major folding event, the transition from the random coil to a near-native state, occurred.

Thermodynamics

The folding thermodynamics is aptly described by a first order-like transition (Hao & Scheraga, 1998) (Figure 5), with the MC folding transition temperature (T_f) estimated to be between 2.1 and 2.3. For high temperatures ($2.0 < T < 3.0$), the data are well-fit by the exponential function typically used to describe two-state folding behavior (Fersht, 1999). This is in general agreement with the folding thermodynamics obtained with lattice models (Shakhnovich, 1997). By matching the backbone dRMS value from NMR measurements (1.48 Å) (Xu *et al.*, 1999), we estimate that room temperature corresponds to $2.1 < T < 2.2$, which is consistent with the general observation that most proteins are marginally stable at room temperature (Creighton, 1992).

Kinetics

Representative folding runs are shown in Figure 6(a)-(e). Although different folding pathways were observed (Figure 7), a fast-folding pathway was found where kinetic traps were avoided. The ensemble of trajectories following this fast-folding pathway was characterized by a definite sequence of events: (1) formation of the inter-helix contacts (event 2 in Figure 7); (2) formation of the

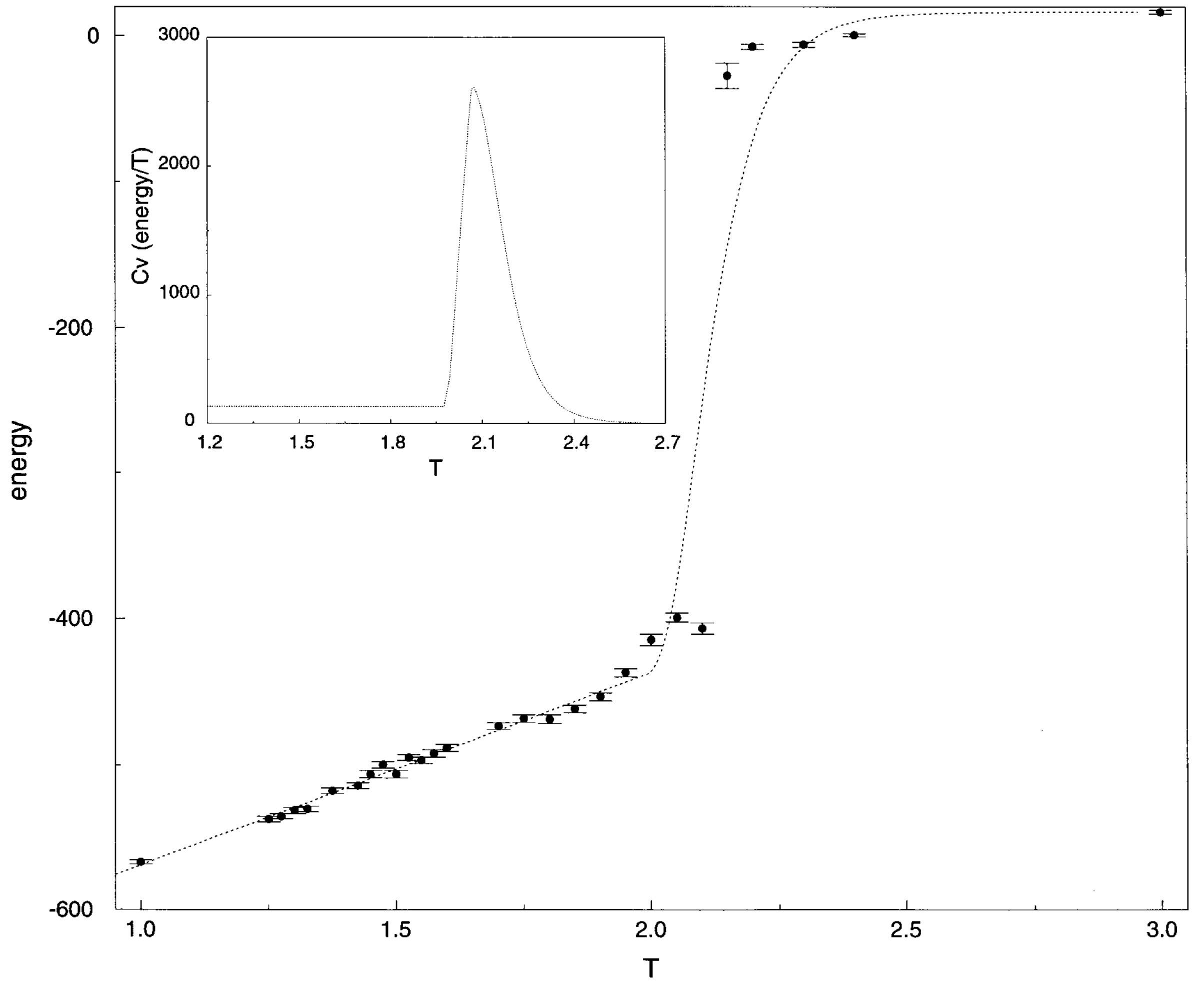

Figure 5. The energy and heat capacity (shown as insert) of crambin as a function of temperature (*T*). The energy data (filled circles) were collected from uncorrelated structures sampled from simulations of 10^8 MC steps initiated from the native state. For $T < 2.0$, the energy data were fitted to a linear function ($R^2 = 0.987$), while for $T \geqslant 2.0$, the exponential function $f(E,T) = E_n + (E_u - E_n)\ \exp(-C/T + D)/(1 + \exp(-C/T + D))$ ($C = 70.82$, $D = 34.16$, $E_n =$ native energy $= -658$, $Eu =$ unfolded energy $= 16.67$; estimated $\sigma \approx 90$) was used. The heat capacity (C_V) was obtained by evaluating $C_V = \mathrm{d}E/\mathrm{d}T$ on the fitted curves. We note that because our simulation does not explicitly model solvent-protein interactions, the computed heat capacity falls to zero for $T > T_f$, in contrast to experimental observations.

two helices (event 3); and (3) formation of the β-sheet (events 4 and 5). This observation can be rationalized from simple topological considerations: the helices are most easily formed when the two ends of the polymer are not constrained by the β-sheet.

At high temperatures ($1.7 < T < 1.8$), at least half of the folding trajectories did not collapse within 10^8 steps, making it clear that collapse is the rate-limiting step (Figures 6(a) and 8(a)). The kinetics appear to be two-state, as evidenced by the narrowly distributed near-native and unfolded populations, rapid collapse events, and no accumulating intermediates. The transiently populated state with energy −150 corresponds to event 2 of the fast-folding pathway (Figure 7) where a subset of the inter-helix and helix contacts have been formed. At middle temperatures ($1.55 < T < 1.675$), this intermediate accumulates for longer times and the collapsed state ensemble broadens (Figure 8(b)). At low temperatures ($T < 1.525$) this effect becomes particularly pronounced, with high energy, collapsed states accumulating for very long times (Figure 8(c)). While collapse is very fast at these temperatures, the persistently broad distribution of low dRMS states (<2 Å) indicates that the compact state is riddled with deep traps. Finally, the median folding time rapidly increased for $T < 1.4$ and $T > 1.8$, and a broad minimum existed at $1.5 < T < 1.65$. More runs are needed to narrow the fastest folding temperature to a smaller range.

Role of helix stability in crambin folding

The zero helix stability at crambin folding temperatures likely explains why the diffusion-collision scenario (Karplus & Weaver, 1976) was not observed as the major fast-folding pathway; in fact, only one out of over 150 runs which folded had the helices forming prior to the inter-helix contacts. It is emphasized that the present move set is not likely to contain any biases against diffusion-collision kinetics. Unlike lattice MC simulations, in which helices can reorient only by partial unfolding, our model allows entire secondary structure elements to move as units. To confirm this point, we performed folding simulations of the two-helix hairpin motif (residues 6-30) in isolation.

In general, if the temperature is sufficiently lowered, we readily observed diffusion-collision kinetics. A representative diffusion-collision folding trajectory of the hairpin at $T = 0.4$ is shown in Figure 9(d). From the ensemble data, it is clear that the diffusion-collision scenario emerges as the dominant kinetic pathway if helix formation is rapid and the helical state is stable (Figure 9(a)-(c)). This requires that the temperature be sufficiently low ($T < 0.8$). Since marginally stable β-hairpin formation is inherently slow (Finkelstein, 1991), at low temperatures where helices fold fast, there is a clear separation of time scales: the helices form first, followed by the hairpin. Given that folding occurs at temperatures where helices are unstable, we must thus rule out diffusion-collision as an important kinetic pathway for this protein

Using a C^{α} off-lattice model with a G$\bar{\text{o}}$ potential, Zhou & Karplus (1999) observed that for a three-helix bundle the "fast track" folding pathway, where no kinetic intermediates are encountered, was the diffusion-collision pathway. In light of our hairpin data, this result is not surprising. The three-helix bundle was being folded at extremely low temperatures: for a bias gap of 1.3 (native interaction $= -1$; non-native interactions $= +0.3$) , the protein was being folded at a temperature that was ≈25 % of the collapse transition temperature (T_f). On our scale, we began to observe crossover to diffusion-collision behavior at $0.8/2.2 \approx 0.36$ of T_f. Although the parameters of the G$\bar{\text{o}}$ potential are slightly different, it is likely that Zhou & Karplus were working at extremely low temperatures where both the helices and the protein were very stable. If we estimate T_f for actual proteins to be roughly 350 K, this implies that the folding simulations by Zhou & Karplus were completed at less than 100 K. It is known experimentally that only the third helix of the bundle is marginally stable (≈30 % helical content) under normal folding conditions (Bai *et al*., 1997). In addition, when a similar three-helix bundle was folded using Langevin dynamics, diffusion-collision behavior was observed at low temperatures, but disappeared as the temperature was raised (G. Berriz & E.I.S., unpublished results). For these reasons, it will be interesting to see if diffusion-collision behavior continues to be observed by Zhou & Karplus at higher temperatures.

Characterization of traps

The kinetic traps observed were either backbone misfolds, where the secondary structure elements were not properly formed, or compact conformations with incorrectly packed side-chains. The first type of backbone misfold resulted when the β-sheet folded incorrectly after the helix-turn-helix moiety was formed. Correction of this misfold required partial or complete unfolding of the β-sheet. At high temperatures ($T > 1.7$), this misfold was metastable, as the β-sheet repeatedly folded and unfolded until the correct topology and packing was achieved (Figure 6(d)). In general, forming the sheet was observed to be the rate-limiting step at low temperatures ($T < 1.5$) (see Figure 8(c)). The other type of backbone misfold occurred when the two helices were not formed properly prior to the formation of the β-sheet. Two pathways (A and B in Figure 7) were available at all temperatures to correct these helix misfolds. Pathway A required the β-sheet to partially or completely unfold. In contrast, pathway B corrected the helix while keeping the sheet intact. This pathway was accelerated with increasing temperature, as the "breathing motion" resulting from greater backbone fluctuations facilitated reinsertion of the helix side-chains (compare, for example, the Q_4 fluctuations in Figure 6(c) and (d)).

The traps resulting from incorrectly packed side-chains (Figure 6(e)) are characterized by low backbone dRMS and correct backbone topology (high Q_is) and were observed only at low temperatures. At higher temperatures, after the major folding event, equilibrium side-chain packing is rapidly achieved. As shown in Figure 6(f), for $T < T^* \approx 1.6$, near-native backbones (dRMS < 1.25 and Q_is > 0.6) obtained from folding runs would not relax further to match the energies attained in runs initiated from the native state. This suggests that T^* may signify a kinetic transition temperature, where ergodicity is broken and a gap emerges between measurements taken from finite but long unfolding and folding runs. At low temperatures, it appears that the major backbone folding event traps side-chains in disordered non-native conformations, which cannot be readily relaxed because of insufficient backbone fluctuations.

Discussion

Compared to molecular dynamics studies of solvated folding (Duan & Kollman, 1998), the approach used in this study is still minimalist. Yet, we have demonstrated that important insights into the folding process, such as the role of side-chain packing, may be obtained by properly combining an all-atom description with simple, atomic resolution energetics. Importantly, our simulation can record a statistically significant number of folding events at atomic resolution for real protein sequences, thereby allowing a relatively direct

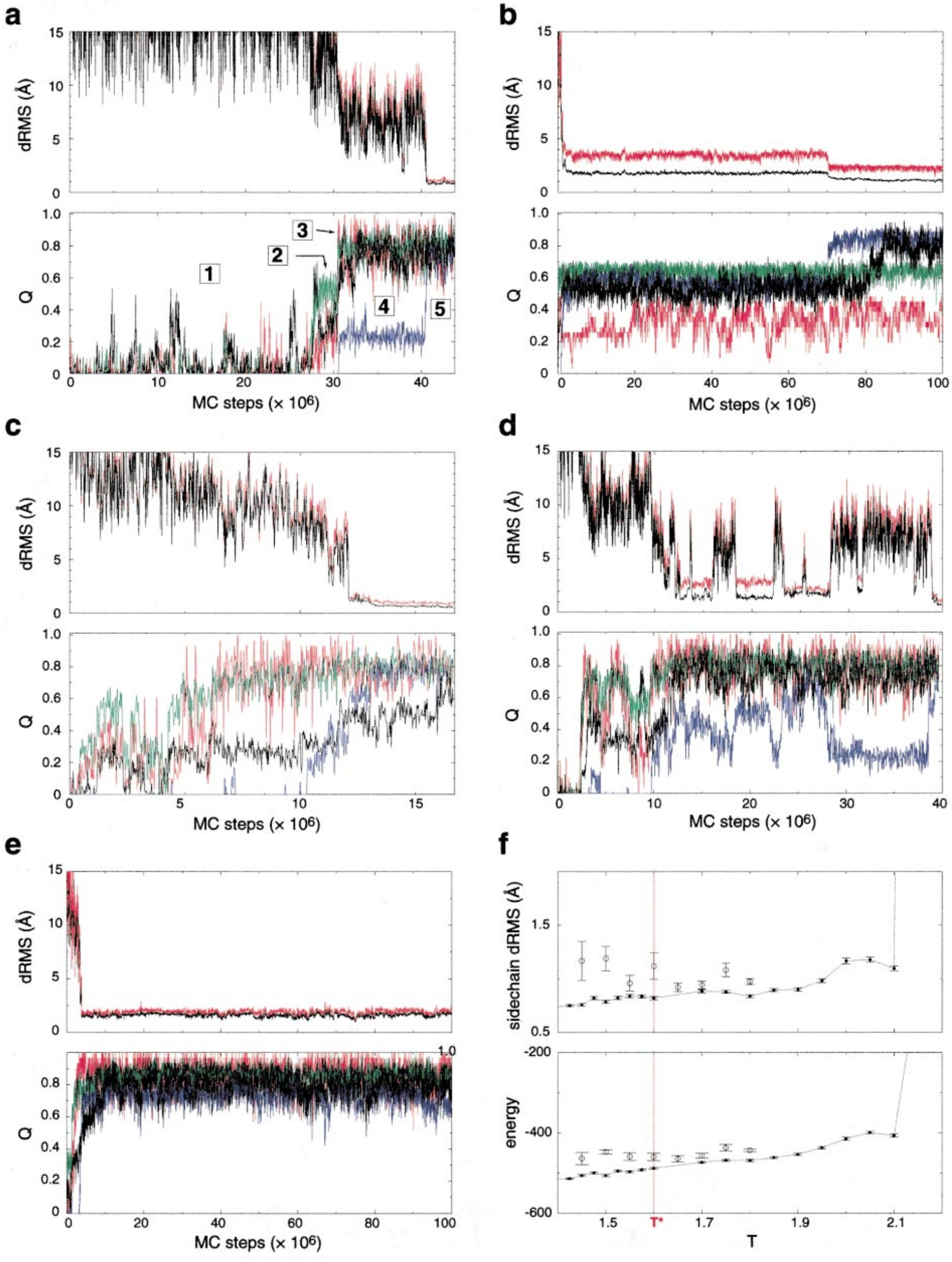

Figure 6 (*legend opposite*)

comparision between simulation and experimental results.

Unfortunately, because of the lack of experimental folding studies on crambin, our results presently cannot be directly verified. We selected crambin for this study because of its small size and its non-trivial α-β structure. We note that as the temperature approaches T_f, the folding kinetics of

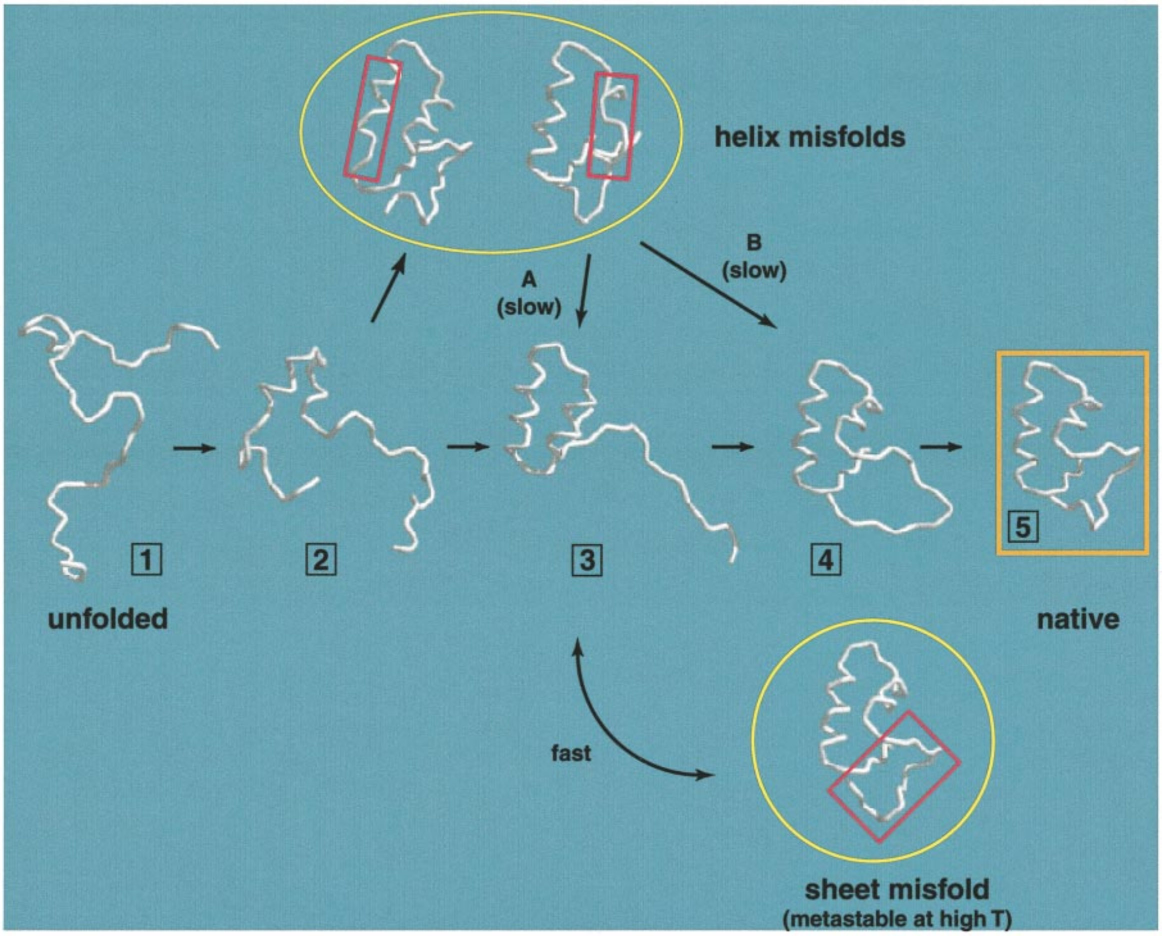

Figure 7. Summary of the folding kinetics. The successive events observed along the fast-folding pathway are marked by boxed numbers. Note that the side-chain packing trap is not indicated on this Figure.

crambin approaches two-state behavior, in agreement with experimental studies of most small single domain proteins (Jackson, 1998). It is plausible that under a Gō potential, where all native interactions are treated identically, the folding properties of crambin we have presented should be consistent with those of similar size and topology.

As noted earlier, helix stabilities were chosen to mimic those observed under physiological conditions. It is experimentally known, however, that the hydrophobic protein crambin is stabilized in a water/alcohol solution (Teeter *et al.*, 1993; Xu *et al.*, 1999). Given that alcohols stabilize helices (Muñoz & Serrano, 1995), it is therefore likely that we have underestimated the stability of helices under conditions of optimal stability for crambin. In light of the fact that Gō energetics are determined primarily by topological constraints and are inherently unable to account for solvent conditions, we can nevertheless ask the following important question: given typical helix stabilities, what folding mech-

Figure 6. Typical folding runs at various temperatures. The upper panel for each run shows backbone (black) and side-chain (red) dRMS as the runs progress. The lower panel tracks the four Q_i values, with the color coding shown in Figure 1. Note that the lengths are different for each run. (a) Collapse-rate limited cooperative folding at high temperature ($T = 1.875$). The secondary structure follows the sequence of events observed in the fast-folding pathway (events 1-5 in Figure 7 are labeled in the Q_i plot). (b) A trajectory ending at a helix 2 misfold at low temperature ($T = 1.25$). (c) Successful folding after encountering a helix 1 misfold at high temperature ($T = 1.775$). This corresponds to pathway B in Figure 7. (d) Successful folding after encountering a β-sheet misfold at high temperature ($T = 1.825$). The β-sheet misfold is metastable at this temperature. (e) A trajectory ending at a low temperature side-chain-packing trap ($T = 1.425$). (f) Broken ergodicity for $T < T^* \approx 1.6$. For each temperature, 52 runs which folded to near-native conformations (dRMS < 1.25 and Q_is > 0.6) were each extended for an additional 25×10^6 steps to allow further relaxation. The average of these extended runs are indicated by open circles. The average values obtained from simulations started from the native state (see Figure 5) are indicated by filled circles.

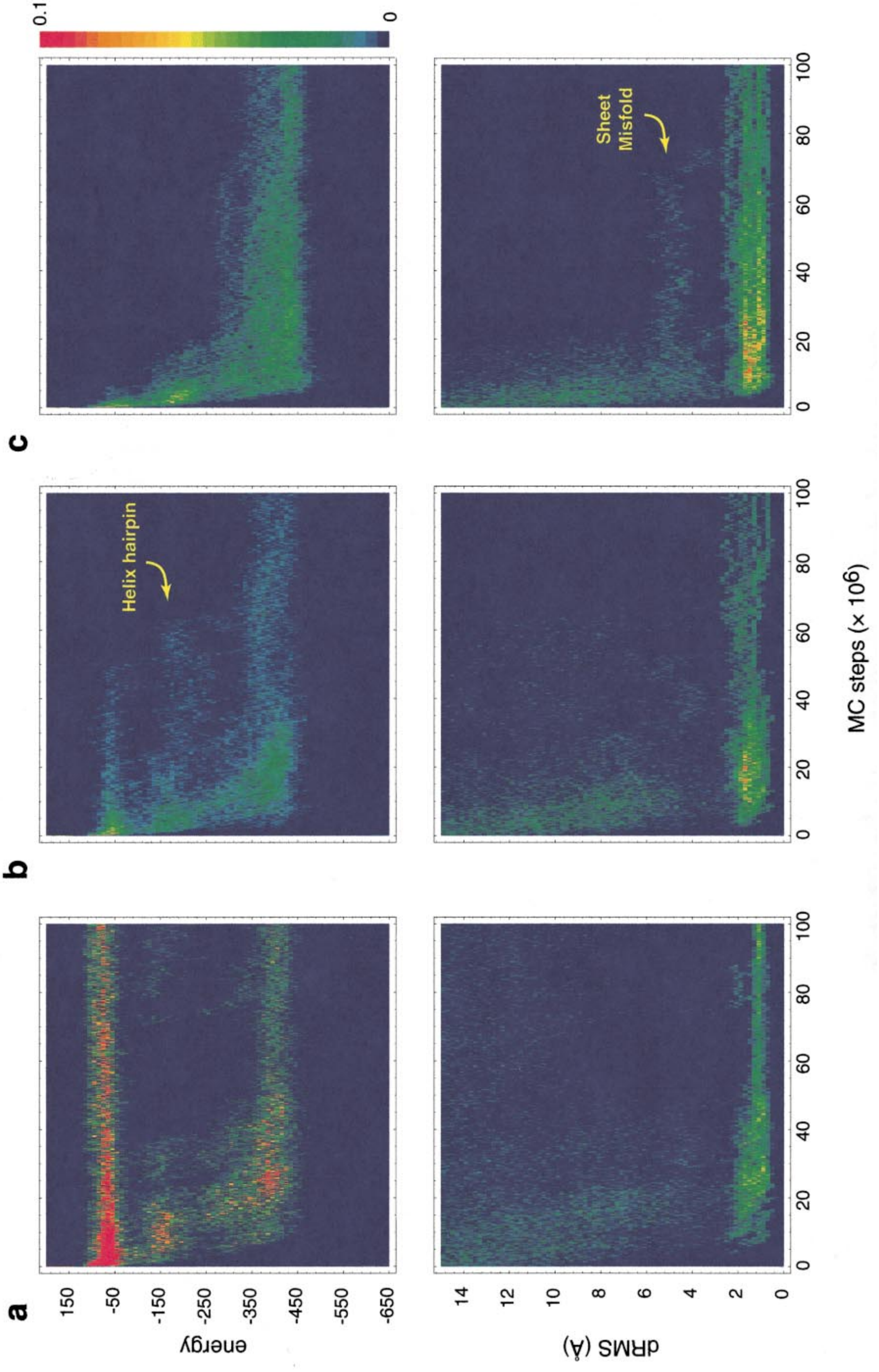

Figure 8 (*legend opposite*)

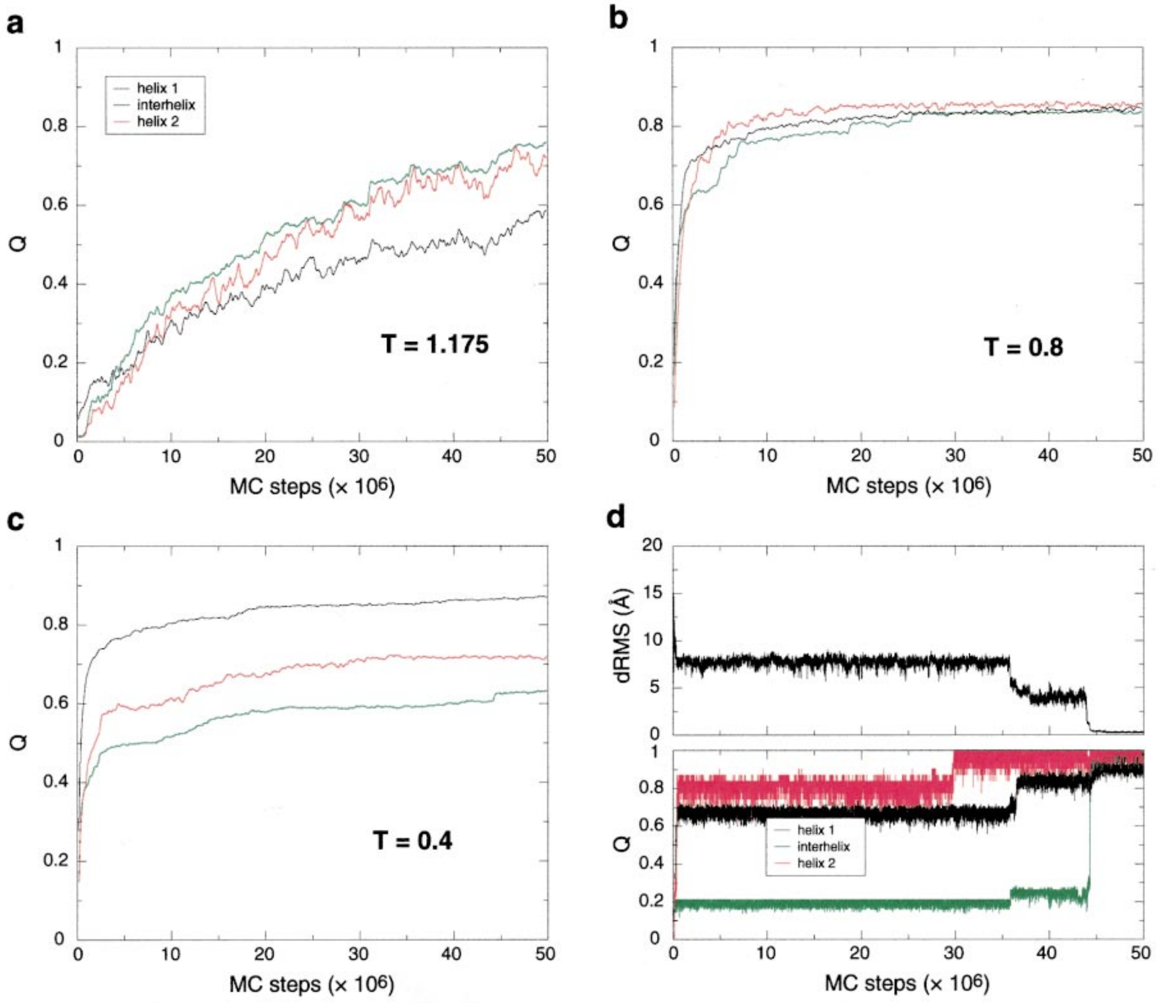

Figure 9. Summary of the hairpin (residues 6-30) formation kinetics. (a)-(c). Ensemble kinetics for various temperatures. The fraction of inter- and intra-helix native contacts is plotted as a function of MC time. The averages were taken over 32 runs of length 5.0×10^7 (d) and (e). A typical folding trajectory at $T = 0.4$. This trajectory exhibits diffusion-collision behavior.

anisms are consistent with a small single-domain protein featuring the same α/β topology such as crambin?

Our data suggest very strongly that a framework-type mechanism (such as the diffusion-collision model) is unnecessary for folding to occur in much less than Levinthal time. In fact, it seems that under conditions (e.g. high simulation *T*) where helix stabilities are low, our results are consistent with the modern viewpoint that nucleation is a major event in folding kinetics (Fersht, 1995; Sosnick *et al.*, 1996; Abkevich *et al.*, 1994b): at temperatures just below T_f, collapse *via* formation of the β-sheet is the major kinetic barrier (Figure 6(a)). A transiently populated, partially collapsed intermediate in which the helix hairpin is formed is seen (Figure 8(a)), and is probably necessary for proper sheet formation. Once the sheet has formed, the chain is compact, and folding proceeds rapidly with concomitant side-chain packing. With the individual helices unstable in isolation at folding temperatures, the hairpin formation did not follow a diffusion-collision pathway. Rather, since the necessary event for hairpin formation is the localiz-

Figure 8. Energy and backbone dRMS kinetics histograms as a function of MC time of all runs at high ($1.7 < T < 1.8$; (a)), middle ($1.5 < T < 1.7$; (b)), low ($T < 1.5$; (c)) temperatures. The color indicates the fraction of runs at a particular energy or dRMS at a given MC time. The color scale goes from blue (0.0) to red (0.1). Because the runs were terminated as they folded, after a run has folded it no longer contributes to the histogram. For this reason, as time progresses and more runs fold, the color of the histograms moves towards blue. At all times, the histogram is normalized by the total number of runs at step 0. The data may be viewed as being obtained by a hypothetical experiment that records the energy and dRMS of all non-folded structures as the simulation progresses. The label hairpin refers to the formation of inter-helix contacts.

ation of inter-helix contacts, the kinetic mechanism is also likely to be a nucleation event, similar to the one theorized for β-hairpins (Finkelstein, 1991).

In addition, structural elements that bring together residues that are closer in position along the chain (such as the helix hairpin) appear to form faster than those which bring together residues distant in sequence position. We believe this is similar to the observation made by Plaxco *et al*. (1998) that folding kinetic rates are correlated with the relative contact order of the native state structure.

Even with a potential strongly biased towards proper folding, the presence of diverse side-chain geometries and excluded volume interactions can lead to the presence of severe kinetic traps, as we observe at very low temperatures. However, our results demonstrate that at reasonable temperatures, side-chains can be successfully packed in a manner consistent with a low dRMS native-like backbone conformation. We believe that the move set we employed contributed to the success of our simulation. Efficient sampling in the compact state, as evidenced by the >10 % acceptance rate even for very low temperatures, resulted from both global and local conformational changes being permitted, with global changes becoming more available with higher temperatures. The qualitative folding behavior is consistent with experimental observations of small single domain proteins, suggesting that the essential features of a polypeptide with all torsional degrees of freedom are captured by this move set.

It is important to note that the helix-coil transition kinetics obtained from our simple model qualitatively showed agreement with experiment and molecular dynamics simulations. Furthermore, whereas current state-of-the-art molecular dynamics simulations focus on extremely short helical segments (penta- and heptapeptides) to study the the helix-coil transition (Hummer *et al*., 2001; Daura *et al*., 1999), we have been able to carry out statistical mechanical analyses on full length helices. The kinetics of early events during folding, such as helix formation, which occur on the timescale of hundreds of nanoseconds (Thompson *et al*., 1997), can thus be investigated with our model provided the folding mechanism is examined from an ensemble viewpoint. Since the median FPT for helix formation is on the order of one million MC steps, this puts the folding of crambin (100 million MC steps) in the ten microsecond range, which is reasonable for a protein of its size. Recent data suggest that the protein G β-hairpin takes on the order of ten million MC steps to fold (J.S. and E.I.S., unpublished results), which places it in the microsecond range to fold in real time, roughly matching the 6 μs rate observed experimentally (Muñoz *et al*., 1997). Furthermore, data on protein G folding indicate a median FPT of one billion MC steps (J.S. and E.I.S., unpublished results). This is one order of magnitude faster than the experimental rate (2-30 ms; Park *et al*., 1999) but the qualitative agreement is promising. It is particularly encouraging to note that the folding rates of secondary structure elements as observed in our simulation are properly separated in timescales. This justifies the use of our simulation to draw qualitative conclusions about kinetic events involving the formation and/or organization of secondary structure elements.

We intend to carry out similar studies on larger single domain proteins for which there are extensive experimental data. The faster folding times we have preliminarily observed for protein G are undoubtedly because of the Gō landscape, which presents an idealistic, perfectly downhill energy landscape. With the development of a sequence-based potential for our model (currently in progress), we expect our correspondence with experimental rates to improve. Taking our results as a proof-of-concept, we believe that detailed investigations of folding pathways may finally be possible using ordinary computational resources.

Methods

Full atom representation

Each non-hydrogen atom present in the crambin crystal structure (Teeter *et al*., 1993) (Brookhaven PDB accession code: 1AB1) was represented by a hard sphere, whose size was given by scaling the relevant VdW radius (*r*) from ref. (Tsai *et al*., 1999) by a factor α (<1). Helix 1 and helix 2 native structures were obtained by extracting residues 6-18 and 23-30, respectively, from the crambin crystal structure.

Move set

A single MC step consisted of a backbone move followed by ten side-chain moves. Each backbone and side-chain move was accepted according to the Metropolis criterion (Metropolis *et al*., 1953). A backbone move consisted of rotating the ϕ-ψ angles of up to three non-proline residues from a randomly selected window of six consecutive residues. A side-chain move consisted of rotating all side-chain torsion angles (χ) of a randomly selected non-proline residue. The size of the backbone and side-chain rotations were obtained from a Gaussian distribution with zero mean and standard deviation 2 and 10 degrees, respectively.

Square well Gō potential

We used an atomic square well potential (McQuarrie, 1976) with the well depths given by Gō energetics (Gō & Abe, 1981). In particular, for two atoms *A* and *B* separated by a distance *R*, the energy ε(*A*,*B*) was calculated according to:

$$\varepsilon(A,B) = \begin{cases} \infty & R < \sigma \\ \Delta(A,B) & \sigma \leqslant R \quad R < \lambda\sigma \\ 0 & R \geqslant \lambda\sigma \end{cases}$$

where $\sigma = \alpha(r_A + r_B)$ is the hard core distance, λ is a scaling factor >1, and $\Delta(A,B) = -1$ if *A* and *B* are in contact in the native conformation and 1 otherwise. The total energy of a conformation was computed as the sum over all pairs:

$$E = \sum_{\text{all pairs}} \varepsilon(A, B)$$

All atom pairs of $i - i + 1$ residues were excluded to eliminate any biases towards local structure, and all backbone-backbone contacts were ignored to eliminate non-specific interactions. The energies of the disulfide bonds were treated no differently from any other contact. We chose $\alpha = 0.75$ because it was the largest value for which the native structure exhibited no steric clashes. Furthermore, with $\alpha = 0.75$, we could not fold crambin with the side-chain torsions held fixed at their native values, suggesting that α was sufficiently large to enforce excluded volume constraints. The selection of small λ values ($\leqslant 1.6$) significantly increased the time of collapse, while large λ values ($\geqslant 2.0$) made side-chain packing more degenerate. We therefore selected $\lambda = 1.8$ in order to balance the two effects. This makes the contact distance for methyl carbon atoms to be 5.08 Å.

Helix-coil transition thermodynamics and kinetics

Thermodynamic data were collected by sampling uncorrelated states observed along long runs of at least 1.5×10^8 MC steps, in order that multiple helix-coil transitions were observed at temperatures near T_f.

Median first passage time data at a given temperature were collected by performing 100 folding runs until the fraction of native contacts Q hit 0.7 or 100×10^6 MC steps had elapsed. For the ensemble kinetic data, 100 folding runs of length 100×10^6 steps were collected, regardless of whether a folding event occurred.

Free energy calculations

Using thermodynamic runs, the average energy as a function of backbone dRMS, $E(r)$, was first measured. The dRMS of the uncorrelated states was next histogrammed in bins of 0.2 Å to compute the probability of observing a particular dRMS, $p(r)$. The entropy as a function of backbone dRMS, $S = \ln\, W(r)$, was obtained by inverting the statistical mechanical relation:

$$p(r) = \frac{W(r)e^{-\beta E(r)}}{Z}$$

where $W(r)$ is the density of states at a given dRMS and Z is the partition function (Ferrenberg & Swendsen, 1988). The partition function was explicitly determined from the $r = 0$ bin by assuming that $W(r = 0) \sim 1$. Finally, the free energy at a temperature T was obtained *via* the identity $G(r) = E(r) - TS(r)$.

Crambin folding kinetics and thermodynamics

Random coils were first generated by unfolding crambin for 3×10^5 steps with only the excluded volume interaction turned on. Both the average energy ($E = 77$, $Q = 0.02$) and average structure ($R_g = 19.5$, backbone $\text{dRMS} = 17.0$) of the random coils indicate that these conformations are completely unfolded and unstructured. Each random coil was then simulated with the square-well G$\bar{\text{o}}$ potential turned on at a particular temperature until it folded or 10^8 MC steps elapsed. Of the 250 runs completed for $T = 1.25$ to 1.875, 165 folded within our observation window of 10^8 steps. For $1.4 \leqslant T \leqslant 1.8$, 135 out of 160 (=84%) runs folded. 2.5×10^6 MC steps approximately took one hour of computation time on a Pentium III 550 Mhz PC.

Acknowledgments

We thank Henk Angerman, Gabriel Berriz, Michael Morrissey, Estelle Pitard, and Juergen Wilder for helpful discussions. This work was supported by NIH grant R01-52126.

References

Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1994a). Free energy landscape for protein folding kinetics: intermediates, traps and multiple pathways in theory and lattice model simulations. *J. Chem. Phys.* **101**, 6052-6062.

Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1994b). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry,* **33**, 10026-10036.

Bai, Y., Karimi, A., Dyson, H. J. & Wright, P. E. (1997). Absence of a stable intermediate on the folding pathway of protein A. *Protein Sci.* **6**, 1449-1457.

Baker, D. (2000). A surprising simplicity to protein folding. *Nature,* **405**, 39-42.

Berriz, G., Gutin, A. M. & Shakhnovich, E. I. (1996). Cooperativity and stability in a langevin model of proteinlike folding. *J. Chem. Phys.* **106**, 9276-9285.

Binder, K. & Heerman, D. W. (1992). *Monte Carlo Simulation in Statistical Physics*, 2nd edit., Springer-Verlag, Berlin.

Bryngelson, J., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Struct. Funct. Genet.* **21**, 167-195.

Bryngelson, J. D. & Wolynes, P. G. (1987). Spin glasses and the statistical mechanics of folding. *Proc. Natl Acad. Sci. USA,* **84**, 7524-7528.

Clementi, C., Nymeyer, H. & Onuchic, J. N. (2000). Topological and energetic factors: what determines the structural details of the transition state ensemble and en-route intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **298**, 937-953.

Creighton, T. E. (1992). *Protein Folding*, 1st edit., W.H. Freeman and Company, New York.

Daggett, V. & Levitt, M. (1992). Molecular dynamics simulations of helix denaturation. *J. Mol. Biol.* **223**, 1121-1138.

Daura, X., van Gunsteren, W. F. & Mark, A. E. (1999). Folding-unfolding thermodynamics of a beta-heptapeptide from equilibrium simulations. *Proteins: Struct. Funct. Genet.* **34**, 269-280.

Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D. & Chan, H. S. (1995). Principles of protein folding - a perspective from simple exact models. *Proteins: Struct. Funct. Genet.* **4**, 561-602.

Dinner, A. R. & Karplus, M. (1999). Is protein unfolding the reverse of protein folding? A lattice simulation analysis. *J. Mol. Biol.* **292**, 403-419.

Duan, Y. & Kollman, P. (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science,* **282**, 740-743.

Ferrenberg, A. M. & Swendsen, R. H. (1988). New Monte Carlo technique for studying phase transitions. *Phys. Rev. Letters,* **61**, 2635-2638.

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.

Fersht, A. (1999). *Structure and Mechanism in Protein Science*, 1st edit., W.H. Freeman and Company, New York.

Finkelstein, A. V. (1991). Rate of beta-structure formation in polypeptides. *Proteins: Struct. Funct. Genet.* **9**, 23-27.

Finkelstein, A. V. (1997). Can protein unfolding simulate protein folding? *Protein Eng.* **10**, 843-845.

Gō, N. & Abe, H. (1981). Noninteracting local-structure model of folding and unfolding transition in globular proteins. I. Formulation. *Biopolymers,* **20**, 991-1011.

Grantcharova, V. P. & Baker, D. (1997). Folding dynamics of the src SH3 domain. *Biochemistry,* **36**, 15685-15692.

Grosberg, A. Y. & Khokhlov, A. R. (1994). *Statistical Physics of Macromolecules*, 1st edit., AIP Press, New York.

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.

Hao, M. & Scheraga, H. A. (1998). Theory of two-state cooperative folding of proteins. *Acc. Chem. Res.* **31**, 433-440.

Hilhorst, H. J. & Deutch, J. M. (1975). Analysis of Monte Carlo results on the kinetics of lattice polymer chains with excluded volume. *J. Chem. Phys.* **63**, 5153-5161.

Honeycutt, J. D. & Thirumalai, D. (1992). The nature of folded states of globular proteins. *Biopolymers,* **32**, 695-709.

Hummer, G., Garcia, A. E. & Garde, S. (2000). Conformational diffusion and helix formation kinetics. *Phys. Rev. Letters,* **85**, 2637-2640.

Hummer, G., Garcia, A. E. & Garde, S. (2001). Helix nucleation kinetics from molecular dynamics simulations in explicit solvent. *Proteins: Struct. Funct. Genet.* **42**, 77-84.

Jackson, S. E. (1998). How do small single-domain proteins fold? *Fold. Des.* **3**, R81-R91.

Jackson, S. E. & Fersht, A. R. (1991). Folding of chymotrypsin inhibitor-2.1. Evidence for a two-state transition. *Biochemistry,* **30**, 10428-10435.

Karplus, M. & Weaver, D. L. (1976). Protein folding dynamics. *Nature,* **260**, 404-406.

Klimov, D. & Thirumalai, D. (1996). A criterion which determines foldability of proteins. *Phys. Rev. Letters,* **76**, 4070-4073.

Kolinski, A. & Skolnick, J. (1994). Monte Carlo simulations of protein folding. II. application to protein A, ROP, and crambin. *Proteins: Struct. Funct. Genet.* **18**, 353-366.

Lazaridis, T. & Karplus, M. (1997). "New view" of protein folding reconciled with the old through multiple unfolding simulations. *Science,* **278**, 1928-1931.

Li, A. & Daggett, V. (1996). Identification and characterization of the unfolding transition state of chymotrypsin inhibitor 2 using molecular dynamics simulations. *J. Mol. Biol.* **257**, 412-429.

McQuarrie, D. A. (1976). *Statistical Mechanics*, 1st edit., Harper Collins, New York.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087-1092.

Muñoz, V. & Serrano, L. (1994). Elucidating the folding problem of helical peptides using empirical parameters. *Nature Struct. Biol.* **1**, 399-409.

Muñoz, V. & Serrano, L. (1995). Elucidating the folding problem of helical peptides using empirical parameters. III. Temperature and pH dependence. *J. Mol. Biol.* **245**, 297-308.

Muñoz, V., Thompson, P. A., Hofrichter, J. & Eaton, W. A. (1997). Folding dynamics and mechanism of beta-hairpin formation. *Nature,* **390**, 196-199.

Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. (1997). Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**, 545-600.

Pande, V. S. & Rokhsar, D. S. (1999). Folding pathway of a lattice model for proteins. *Proc. Natl Acad. Sci. USA,* **96**, 1273-1278.

Pande, V. S., Grosberg, A. Y., Rokshar, D. & Tanaka, T. (1998). Pathways for protein folding: is a "new view" needed? *Curr. Opin. Struct. Biol.* **8**, 68-79.

Pande, V. S., Grosberg, A. Y. & Tanaka, T. (2000). Heteropolymer freezing and design: towards physical models of protein folding. *Rev. Mod. Phys.* **72**, 259-314.

Park, S. H., Shastry, M. C. R. & Roder, H. (1999). Folding dynamics of the B1 domain of protein G explored by ultrarapid mixing. *Nature Struct. Biol.* **6**, 943-947.

Plaxco, K. W., Simons, K. T. & Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985-994.

Ramanathan, S. & Shakhnovich, E. I. (1994). Statistical mechanics of proteins with "evolutionary selected" sequences. *Phys. Rev. E,* **50**, 1303-1312.

Rey, A. & Skolnick, J. (1991). Comparison of lattice Monte Carlo dynamics and Brownian dynamics folding pathways of α-helical hairpins. *Chem. Phys.* **158**, 199-219.

Richards, F. M. & Lim, W. A. (1994). An analysis of packing in the protein folding problem. *Quat. Rev. Biophys.* **26**, 423-498.

Sali, A., Shakhnovich, E. I. & Karplus, M. (1994). How does a protein fold? *Nature,* **369**, 248-251.

Scholtz, J. M. & Baldwin, R. L. (1992). The mechanism of α-helix formation by peptides. *Annu. Rev. Biophys. Biomol. Struct.* **21**, 95-118.

Shakhnovich, E. I. (1997). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29-40.

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.

Sosnick, T. R., Mayne, L. & Englander, S. W. (1996). Molecular collapse: the rate-limiting step in two-state cytochrome c folding. *Proteins: Struct. Funct. Genet.* **24**, 413-426.

Teeter, M. M., Roe, S. M. & Heo, N. H. (1993). Atomic resolution crystal structure of the hydrophobic protein crambin at 130 K. *J. Mol. Biol.* **230**, 292-311.

Thompson, P. A., Eaton, W. A. & Hofrichter, J. (1997). Laser temperature jump study of the helix-coil kinetics of an alanine peptide interpreted with a "kinetic zipper" model. *Biochemistry,* **36**, 9200-9210.

Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. (1999). The packing density in proteins: standard radii and volumes. *J. Mol. Biol.* **290**, 253-266.

Verdier, P. H. (1973). Monte Carlo studies of lattice-model polymer chains. 3. Relaxation of Rouse coordinates. *J. Chem. Phys.* **59**, 6119.

Xu, Y., Wu, J., Gorenstein, D. & Braun, W. (1999). Automated 2D NOESY assignment and structure calculation of crambin (S22/I25) with the self-correcting distance geometry based NOAH/DIAMOD programs. *J. Magn. Reson.* **136**, 76-85.

Young, W. S. & Brooks, C. L., III (1995). A microscopic view of helix propagation: N and C-terminal helix growth in alanine helices. *J. Mol. Biol.* **259**, 560-572.

Zhou, Y. & Karplus, M. (1999). Interpreting the folding kinetics of helical proteins. *Nature*, **401**, 400-403.

Zimm, B. H. & Bragg, J. K. (1959). Theory of the phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys.* **11**, 526-535.

Edited by A. R. Fersht

(Received 22 November 2000; received in revised form 28 February 2001; accepted 5 March 2001)