Folding pathway of a lattice model for protein folding

Vijay S. Pande[1] and Daniel S. Rokhsar[1,2]

The folding of a protein-like heteropolymer is studied by direct simulation of a lattice model that folds rapidly to a well-defined "native" structure. The details of each molecular folding event depend on the random initial conformation as well as the random thermal fluctuations of the polymer. By analysing the statistical properties of hundreds of folding events, a classical folding "pathway" for such a polymer is found which includes partially folded, on-pathway intermediates that are shown to be metastable equilibrium states of the polymer. The folding reaction has a well-defined ensemble of transition state conformations that are characterized by common core structures. (January 4, 1998; Revised March 1, 1998)

Small proteins typically fold rapidly and reliably to a unique native state from any one of a vast number of initial unfolded conformations. How does this occur? The rapidity of folding, the complexity of protein structure, and the fact that each molecule takes a microscopically different path to the folded state makes this a difficult problem for both experiment (which typically provides ensemble averaged information with limited temporal and spatial resolution) and all-atom simulations (which can only examine protein motions for nanoseconds rather than the milliseconds needed for folding) [1].

If there are general principles that describe protein folding, then one might expect them to apply to simplified models for protein-like heteropolymers as well. For this reason, models for polymer folding on a lattice have been vigorously pursued [2–9]. Although lattice models often omit features that are critical for understanding protein function, they are "protein-like" in the sense that they fold to a unique native structure from an astronomically large number of possible initial conformations, and do so rapidly, reproducibly, and reversibly. The advantages of these models are two-fold. First, the thermodynamic driving force for folding is an explicit part of the model, so that the origin of the stabilizing forces can be separated from the problem of folding mechanism. Second, it is straightforward to study a large collection of folding events from start to finish by direct simulation, with complete access to the structural details of every conformation that the polymer samples, without any "dead time." The challenge is then extracting a useful description of the folding process from the abundance of detail provided by these numerical experiments.

Here we study the folding of a 48-mer on a three-dimensional cubic lattice for which the inter-residue interactions have been chosen to stabilize a pre-selected "native" conformation, shown in the last frame of Fig. 1 (see Methods). This heteropolymer exhibits a cooperative, two-state transition with a midpoint temperature of $T_m = 0.74 \pm 0.01$ (in units of the interaction strength, see Methods) from an unfolded phase U with no persistent structure to a folded phase N consisting of the native conformation and small fluctuations about it. Using Monte Carlo dynamics (see Methods) and a new "fluctuation smear" analysis technique (see below), we examine the sequence of conformations encountered when an initially unfolded polymer is suddenly quenched to a temperature at which the folded state is thermodynamically stable (*i.e.*, below T_m). Each complete simulation – a "folding event" – begins from a different initial unfolded conformation, and proceeds until the native conformation is reached. We follow hundreds of independent folding events, and extract their common features – the "folding pathway."

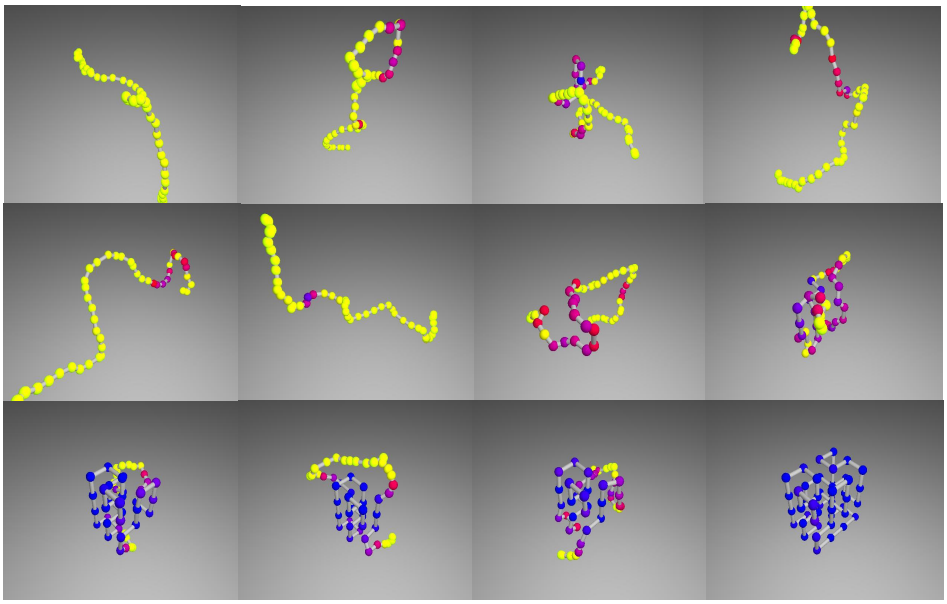

FIG. 1: Snapshots of a folding event. The average trace of the polymer backbone during the course of a single folding event is shown at intervals of 6×10^4 Monte Carlo steps. The average is taken over a time window of 10^4 steps. (While the individual conformations lie on the lattice, the average positions need not do so.) The short-time fluctuations of the polymer conformation are encoded by the color of the residue: blue indicates parts of the polymer that are relatively static over the time window (positional variance $\langle|\delta\mathbf{r}_i|^2\rangle < 0.5$), while yellow indicates sections whose position fluctuates strongly ($\langle|\delta\mathbf{r}_i|^2\rangle > 3$); the fluctuations of residues colored red lie in between. A nucleation event [13, 14] is evident at $t \approx 4.9 \times 10^5$ steps. The final frame shows the native state conformation.

[1]Department of Physics, University of California at Berkeley, Berkeley, California 94720-7300
[2]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

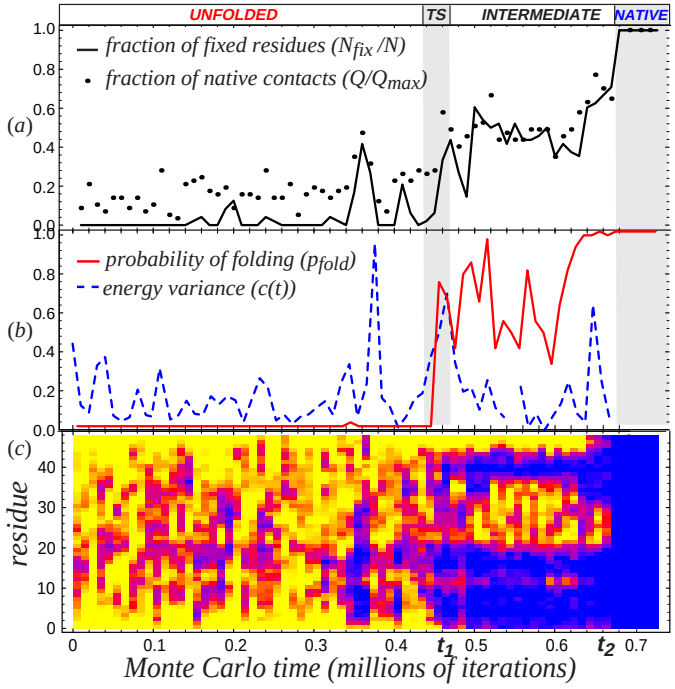

FIG. 2: Qualitative and quantative probes of a folding event. **(a)** shows the number of native contacts Q and the number of "fixed" (blue) residues $N_{\rm fix}$ along the folding trajectory. **(b)** shows the folding probability $p_{\rm fold}$ (see Methods) and the energy variance $\langle(\delta E)^2\rangle$ of the conformations between $t \pm 1.5 \times 10^4$. **(c)** shows the positional variance $\langle|\delta \mathbf{r}_i|^2\rangle$ *vs.* time on the horizontal axis and position along the chain on the vertical axis, using the same color coding as in Fig. 1. Note that the static (*i.e.,* blue) residues are clustered in patches for $t_1 < t < t_2$, representing the persistent core of the intermediate; for $t > t_2$ all residues become fixed in their native positions, and the polymer is folded.

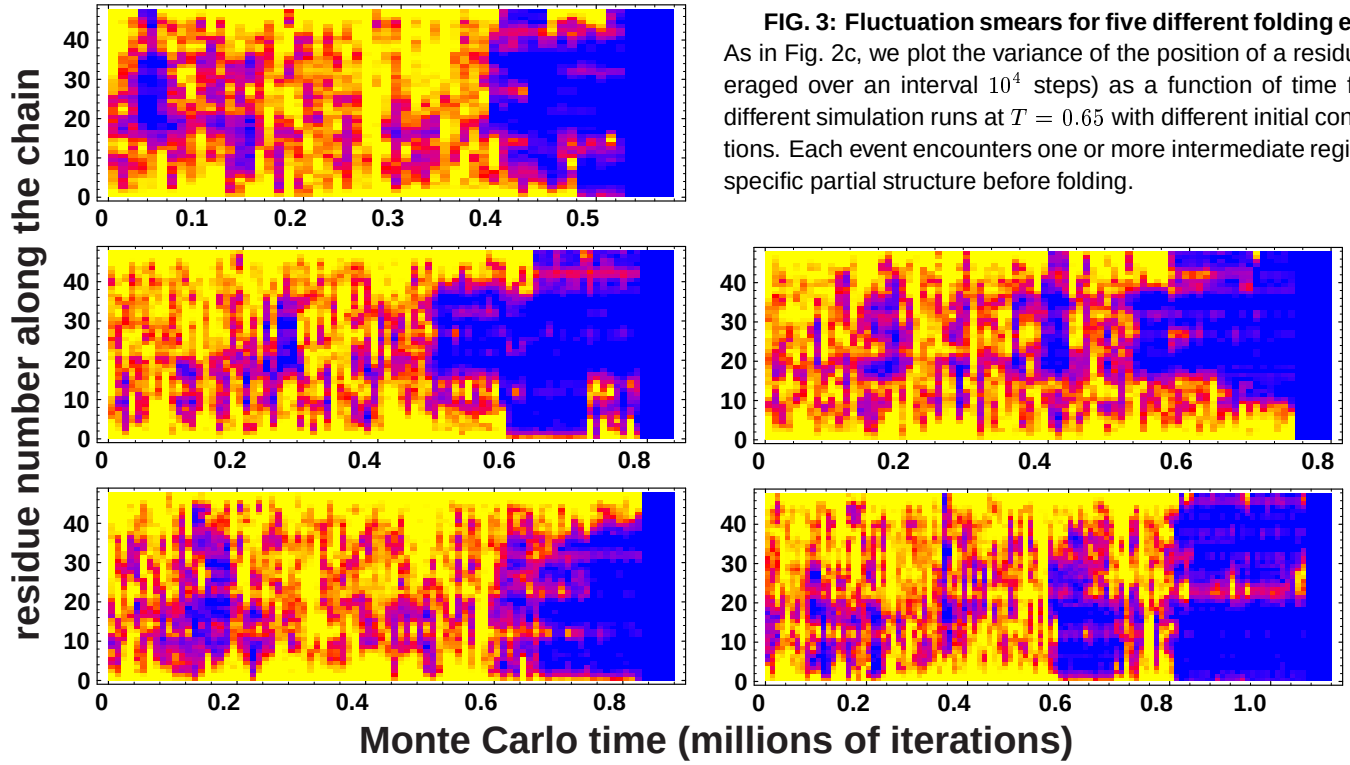

FIG. 3: Fluctuation smears for five different folding events. As in Fig. 2c, we plot the variance of the position of a residue (averaged over an interval 10^4 steps) as a function of time for five different simulation runs at $T = 0.65$ with different initial configurations. Each event encounters one or more intermediate regimes of specific partial structure before folding.

Our analysis focusses on the nature of the conformations that the polymer samples as it approaches its stable native state, and the characterization of the transition state and any intermediates it may pass through on the way. What sets our approach apart from earlier work is our attention to the local and short-time fluctuations of the polymer as it folds (visualized by smear analysis), and the characterization of an ensemble of folding trajectories using hundreds of independent folding events under identical conditions. We find that our model proteins fold *via* partially folded "on-pathway" intermediates, with a well-defined transition state ensemble that we characterize statistically.

Anatomy of a folding event

Fig. 1 illustrates twelve frames at 6×10^4 time step intervals from a single folding event at temperature $T = 0.65$, where each frame represents the average position of each residue during a time win-

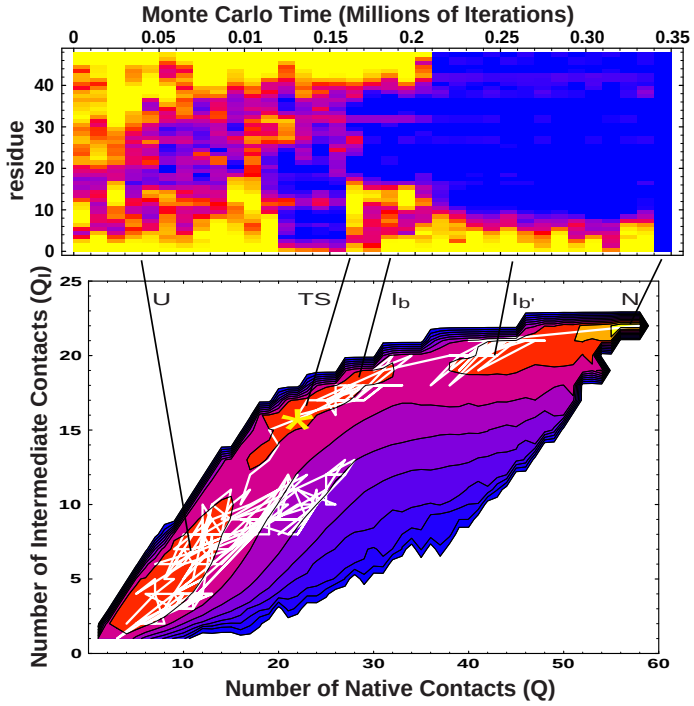

FIG. 4: Kinetic intermediates and free energy minima. (a) Fluctuation smear. **(b)** Contours of the free energy surface $G(Q, Q_{I-a})$ projected onto the (Q, Q_{I-a}) plane are shown superimposed on the trajectory of the folding event shown in (a). The free energy surface was computed using umbrella sampling during a long equilibrium run (see Methods); $Q_{I-b} \equiv \sum_{ij} C_{ij} C_{ij}^{\mathrm{I-b}}$ is the number of contacts a conformation shares with a representative conformation $C_{ij}^{\mathrm{I-b}}$ chosen from the intermediate of (a). The three distinct free energy minima correspond to the unfolded, intermediate, and native states; the native state is the most stable of the three; a later intermediate is also evident. (The free energy minima corresponding to the other metastable intermediates are not visible in this projection onto the (Q, Q_{I-b}) plane.) The trajectory shows that the polymer fluctuates in the unfolded state before rapidly jumping to the intermediate; from the intermediate complete folding proceeds relatively rapidly. The transition state for this trajectory (the conformation where p_{fold} first crosses 1/2) is indicated by an asterisk.

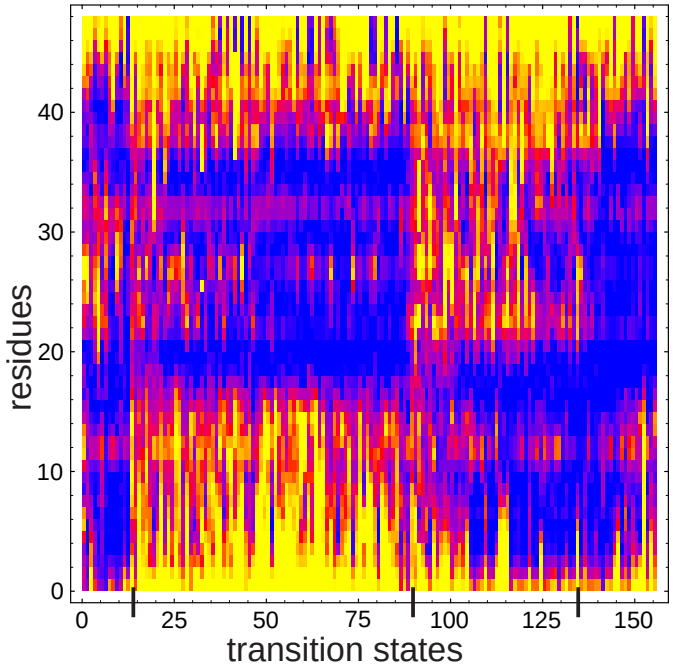

FIG. 5: Characterization of the Transition State Ensemble. The variances of transition state conformations from 156 independent folding runs at $T = 0.65$ are displayed using the color coding scheme of Fig. 1. Conformations have been sorted to exhibit the natural partitioning of these conformations into distinct transition state "classes." Superimposed on the common structure within each class are various optional structures; these contribute to the entropy of the transition state.

dow of 10^4 steps. (A movie of this folding event is available at *http://hubbell.berkeley.edu/nsb.html*.) The residues are color-coded to display the fluctuation in their position during this time interval, as indicated in the caption. The first eight frames show the polymer in a fluctuating, unfolded state. The ninth, tenth, and eleventh frames show persistent partial structure (indicated by the blue structured region). In the final frame, the native state is abruptly reached.

The same folding event is examined in more detail in Fig. 2. Fig. 2c displays the fluctuation in position of the i-th residue, using the same color code as in Fig. 1. This fluctuation "smear" (Fig. 2c) clearly exhibits three distinct regimes separated by sharp transitions. The polymer begins in an initial, highly fluctuating unfolded state that persists up to $t_1 \approx 4.9 \times 10^5$. It then abruptly switches to a series of conformations with common partial structure, as indicated by the emergence of persistent blue streaks. These conformations possess a *specific* set of non-fluctuating ("fixed") residues with high probability. The nature of these contacts is clearly shown in Fig. 1, panels 9-11: the contiguous blue regions along the chain represent a fixed "core" of contacts – both "secondary" (*i.e.*, nearby along the chain) and "tertiary" – while the intervening red region (residues 18-39) represent a fluctuating internal "loop."

At time $t_2 \approx 6.8 \times 10^5$, there is a rapid transition from the partially folded intermediate state to the completely folded native conformation. This sharp transition corresponds to the rapid absorption of the fluctuating internal loop. At the temperature of the simulation, the unfolding rate from the native state is quite slow, and the polymer remains folded for typically 5×10^7 iterations (data not shown). The fluctuation smear analysis shows that the time evolution of the folding polymer divides into discrete regimes that we characterize below as the unfolded state and metastable intermediates.

We have studied hundreds of folding events at each of several temperatures, but here focus on $T = 0.65$, at which folding typically takes about a million time steps. In these events, the polymer initially remains in the unfolded state for several hundred thousand iterations before undergoing a transition to an intermediate with persistent partial structure. After lingering in an intermediate for a comparable time, the polymer either folds, returns to the unfolded state, or, more rarely, makes a sharp transition to a distinct intermediate regime. The polymer may encounter several intermediate states before finally folding to the native conformation. (Some trajectories also linger in nearly folded conformations with $Q/Q_{\max} > 0.9$ immediately prior to complete folding; we regard these conformations as small fluctuations from the native state.) Fig. 3 illustrates this behavior with fluctuation smears for five more folding events at the same temperature as Figs. 1 and 2. This behavior is common to all simulated folding events we studied.

The transitions between the unfolded and intermediate regimes, and between the intermediates and the native state, are sharp and highly cooperative, as shown in Figs. 2 and 3. Although here we will often refer to a single event in detail, it must be emphasized that our general results derive from the analysis of hundreds of independent folding trajectories. We characterize the folding pathway of our model protein by (a) identifying the unfolded state and intermediate(s) that are sampled by the folding polymer, and (b) determining the dominant transition state for folding, which for our model system is the transition from the unfolded state to a partially folded intermediate.

The unfolded state

Beginning from an initial unfolded conformation, the polymer rapidly reaches a (metastable) thermal equilibrium. During this first regime ($t < t_1$ in Figs. 1 and 2) the polymer samples conformations with several native contacts (typically $Q/Q_{\max} = 0.18 \pm 0.09$), but these contacts are fluctuating: *specific* contacts are not preserved from conformation to conformation, since hardly any residues are stationary ($N_{\rm fix}/N \approx 0$), as shown in Fig. 2b. Here $Q(t)$ is the number of native contacts found in conformations at time t, and $N_{\rm fix}(t)$ is the number of residues with low positional variance (*i.e.,* blue) in these conformations. ($Q_{\max} = 57$ is the number of contacts in the native conformation and $N = 48$ is the total length of the chain.)

This rapidly interconverting set of conformations with small Q is found in all folding events, and defines the unfolded "phase" of the polymer. Since below T_m the thermodynamically stable state of the polymer is the native state, this unfolded state must be regarded as a metastable or "supercooled" phase. The existence of flickering native contacts in the unfolded state of our protein-like heteropolymer is consistent with the observations of transient structures in the unfolded states of proteins [15–17].

Partially folded intermediates

In Fig. 2, near time t_1, persistent contacts suddenly form, and $N_{\rm fix}/N$ rises rapidly to become comparable to $Q/Q_{\max}$. There is only a gradual increase in $Q/Q_{\max}$ at this time. During the interval $t_1 < t < t_2$ both of these measures of folding progress remain nearly constant. Similar intervals of persistent partial structure appear in all folding events we have examined. These sharply-defined intermediate regimes define collections of rapidly interconverting conformations that comprise the intermediate phases of the polymer. These "kinetic" intermediates are transient, and do not accumulate. Note that the presence of kinetic intermediates is consistent with a cooperative equilibrium two-state transition, in the sense that only the unfolded and native states are ever appreciably populated in equilibrium.

We find that the intermediate regimes found in the folding events we studied can be classified into four distinct classes, each with its own well-defined, persistent partial structure. (These classes are discussed further below.) It is not uncommon for the polymer to encounter several intermediates in a single folding event. When this occurs, the intermediates are typically separated by a return to the unfolded state (see Fig. 3). We also occasionally observe rare direct transitions between intermediates that involve either the formation or dissolution of a discrete unit of partial structure. Since we never see direct $U \rightarrow N$ transitions, we conclude that for this system folding must proceed through an intermediate. In this sense, the intermediates we find are obligatory steps in the folding pathway of this polymer, and can be classified as "on-pathway."

The sudden, discrete changes in the character of the conformational fluctuations of the polymer found in all folding events (*e.g.,* at t_1 and t_2 in Fig. 2) are naturally interpreted as highly cooperative transitions involving metastable phases. This cooperative character is further demonstrated by considering the time-dependent heat capacity $c(t)$ (see Methods). Clear peaks in $c(t)$ are seen at both boundaries (*i.e.,* t_1 and t_2) in Fig. 2b, indicating sudden releases of heat from the polymer to the thermal bath that reflect the latent heats released at cooperative transitions. (The large $c(t)$ peak seen at $t \approx 3.5 \times 10^5$ steps corresponds to the brief formation of a partially folded structure that quickly dissolved, which appears as a short blue patch in Fig. 2c.)

Further evidence that the unfolded and partially folded regimes can be identified with metastable phases is provided by *equilibrium* Monte Carlo umbrella sampling (see Methods) of the free energy $F(Q, Q_I; T)$, *i.e.,* the free energy as a function of the number of native contacts, Q, and the number of contacts Q_I that are found in common with a representative intermediate conformation. Fig. 4b exhibits *three* distinct local free energy minima corresponding to the unfolded, intermediate, and native phases, with barriers between them. Their relative stability depends on temperature; for $T < T_m$ the native state has lowest free energy. (The local free energy minima corresponding to the other intermediates are not visible with this choice of thermodynamic coordinates.) The intermediate states for this polymer are never global free energy minima, since either the unfolded and/or native states are always lower in free energy at all temperatures.

Determining the transition state ensemble

Each folding event samples a different microscopic series of conformations on its way to the final folded state. The transition state for folding is therefore an *ensemble* of conformations, and must be characteristized statistically [18]. Here we implement a new computational method [19] that allows an unambiguous determination of the transition state for a complex reaction such as folding.

To determine whether a given conformation C is a member of the transition state ensemble, we appeal to the defining feature of a transition state as an unstable species that is equally likely to continue to the completion of the reaction (*i.e.,* to fold) or to revert to the unreacted state (*i.e.,* to unfold). To measure the progress of the reaction we calculate the "folding probability" $p_{\rm fold}(C)$ [19], which is the probability that a new simulation that starts from conformation C reaches the native state without first unfolding (see Methods). In a reaction without intermediates, the transition state lies at $p_{\rm fold} = 1/2$; for more complex reactions, $p_{\rm fold} \approx 1/2$ determines the dominant transition state. (Note that conformations with $p_{\rm fold} = 1/2$ may be encountered several times in a given event.) The folding probability method permits the determination of the transition state without requiring any assumptions regarding the nature of the reaction coordinate or free energy surface.

Fig. 2b shows the folding probability as a function of time for the folding event shown in Fig. 1. Throughout the initial unfolded regime, $p_{\rm fold}$ remains zero. A rapid increase in $p_{\rm fold}$ through $1/2$ coincides with the transition between the unfolded and intermediate regimes, when partial, native-like structure first appears. This is shown more dramatically in Fig. 4b, which exhibits a folding trajectory superimposed on the calculated free energy surface in the (Q, Q_I) plane, where Q_I measures the number of contacts in common with a representative intermediate conformation. (The fluctuation smear for this event is shown in Fig. 4a.) The conformation at which $p_{\rm fold}$ first exceeds 1/2 is indicated by an asterisk. We find no apparent precursor to the abrupt increase in $p_{\rm fold}$ in any of the diagnostic quantities we have measured on time-scales of order 10^4 steps, indicating that the change occurs rapidly, *i.e.,* over the course of a few hundred conformations, and does not occur by a gradual accumulation of native structure. Similar abrupt behavior occurs in every folding run we examined.

The transition state ensemble can be constructed operationally by collecting the conformations at which $p_{\rm fold}$ first enters the range $0.4 < p_{\rm fold} < 0.6$ from many independent folding events. Fig. 5 displays the positional variance of the transition states of 156 independent folding events at $T = 0.65$ with the same color coding as used earlier. (In 134 other events, $p_{\rm fold}$ jumps from below 0.4 to above 0.6 faster than our $p_{\rm fold}$ sampling interval of 10^4 steps; the first conformations encountered with $p_{\rm fold} > 0.6$ in these runs are similar to those shown in Fig. 5, data not shown.)

Transition state classes

In our simulations, typical transition state conformations contain a compact, native-like core (shown in blue in Fig. 5), surrounded by fluctuating polymer loops or dangling ends (shown in yellow/red). While the transition state conformations vary from one folding event to another, Fig. 5 shows that they can be grouped into four distinct classes: (a) core contacts involving residues 2-22 and 33-40, with an internal loop and one short dangling end, (b) core involving residues 18-37, with two dangling ends, (c) core with residues 2-18, with one long dangling end, and (d) a combination of the cores in (b) and (c). The transition state of the trajectory shown in Figs. 1 and 2 is in class (a); the transition state for the event in Fig. 4 is in class (b).

A comparison of Figs. 5 and 2 shows that transition state class (a) is simply a less ordered version [20] of the intermediate found in the folding event shown in Figs. 1 and 2. In general we find that each transition state class is associated with its own intermediate phase, so that Fig. 5 can also be used to characterize the metastable partially structured phases of the polymer. Such a pairing of transition state and intermediate is consistent with the "post-critical nucleus" approach of ref. [7]. All four intermediates can be reached directly from the unfolded state *via* their corresponding transition state classes, and all four can fold directly to the native state. Direct transitions between intermediates are rarely found in the folding events we have studied, and only occur between intermediates that share common partial structure: while the transitions $I_b \rightarrow I_d$ and $I_c \rightarrow I_d$ do occur, we have never observed intermediate-intermediate transitions involving I_a. The relationship between transition state classes and intermediates can also be seen by comparing the fluctuation smears shown in Fig. 3 with Fig. 5.

Kinetic intermediates and metastable states

Many single-domain proteins – apomyoglobin [32, 33], barnase [39], interleukin 1β [41], *etc.* – fold via one or more kinetic intermediates, which can sometimes be stabilized under acid or denaturing conditions as equilibrium states []. These intermediates appear to possess partial native structure; for example, in the acid state of apomyoglobin, the AGH-helices form a structured region [33]. Recent theories of protein folding, however, have emphasized the role of misfolded intermediates that impede folding by trapping the protein in a locally non-native structure [8, 12, 21, 22]. This scenario is in keeping with the fact that many small proteins, such as chymotrypsin inhibitor II [36], lambda repressor [37], and the cold shock protein Csp [38], fold rapidly without any detectable intermediates.

Here, we have studied a model polymer that folds to a unique native structure through a folding pathway that is characterized by sudden, cooperative transitions between several distinct, partially folded intermediates. These intermediates are obligatory steps in the folding pathway. They are also thermodynamically metastable "phases" of the polymer under native conditions, in the sense that they correspond to local free energy minima. Thus these intermediates represent both rare equilibrium fluctuations from the native state as well as transient kinetic species in the folding pathway.

The intermediates we observe are naturally identified with the "partially unfolded states" (PUFs) found using native state hydrogen exchange in cytochrome C [25] and RNase H [26]. These PUFs must be thought of as metastable thermodynamic states of the protein, since if there were no free energy barrier between partially unfolded states and the native state, one would find a *continuous* distribution of hydrogen exchange protection dependences rather than the discrete groupings that are observed.

The identity of metastable equilibrium states and transient kinetic intermediates that we find for our model polymer parallels the situation in RNase H, in which the kinetic folding intermediate has been shown [27] to resemble both a rare partially unfolded state observed in equilibrium under native conditions and the equilibrium molten globule [28–30] stabilized under acid conditions. The intermediates we find also have the qualities of the molten globule-like state that is stabilized under "denaturing" conditions in lattice models [31]. The connection between kinetic in-

termediates, partially unfolded metastable states, and equilibrium molten globules that we find in our model supports the hypothesis that molten globules represent stabilized versions of intermediate states along the folding pathway [32–34].

The intermediates we find are qualitatively different from those observed in other theoretical studies. Previous simulations of other protein-like heteropolymers, using different models and/or polymer lengths, found intermediates which are either misfolded [8, 21, 22, 24] or correspond to the folding of domains [23]. In their studies of 36-mers using a sequence model, Shaknovich *et al.* [21, 22] find intermediates that are characterized as off-pathway traps that lead to slower folding than in the absence of intermediate. For some longer chains, multidomain behavior was observed [23] in which partially folded intermediates can be observed as the equilibrium state under some conditions. Thirumalai *et al.* [24] have found misfolded intermediates in an off-lattice model. Based on their studies of 27-mers, Onuchic *et al.* have also stressed that intermediates appear as off-pathway traps that slow folding [8].

By contrast, the partially folded, metastable states we find have no non-native structure, and are "on-pathway" intermediates in the sense that folding can proceed directly to the native state. Our model shows that intermediates are not necessarily a consequence of trapping in a misfolded conformation. We stress that there is no contradiction between our results and those found using other models and polymer lengths; the fact that our simulations show partially folded on-pathway and others find misfolded traps in theirs is consistent with the diverse folding pathways of different proteins.

Transition states in theory and experiment

Rather than a single enthalpically unfavorable conformation, the transition state for protein folding is necessarily an *ensemble* of conformations, and can only be characterized statistically. Previous theories of protein folding based on simplified models relied on a variety of criteria to infer this ensemble, which led to proposed transition states with different characteristics: native-like conformations [5], a unique, specific nucleus [7], many delocalized nuclei [8], or a potentially broad ensemble [12]. These varied proposals may reflect an inherent diversity in protein folding pathways themselves [12].

Here we have studied the folding pathway of a longer polymer which, unlike those studied in refs. [8, 21–24], must pass through one or more partially folded intermediates before folding. For our 48-mer, we have determined the transition state ensemble for folding by directly identifying those conformations in each folding event that are equally likely to fold or unfold. For the folding of this polymer, the major transition state occurs between the unfolded and intermediate states. The transition state of a typical folding event consists of a well-defined core of residues that are surrounded by fluctuating polymer loops or dangling ends. The conformations in the transition state ensemble can be naturally partitioned into a few distinct groups of related conformations, or "classes," each containing a specific set of core residues, with relatively minor variations of these residues within a class. These classes are closely related to the "specific nucleus" of Shakhnovich *et al.* [7]. The formation of this core structure represents a nucleation event [7, 13, 14, 35] that is the critical event in the folding pathway of this polymer.

The transition state for our polymer is dominated by the two classes (b) and (c), with (a) and (d) as minor components. The existence of a few distinct transition state classes (and associated intermediates) is typical of the lattice heteropolymers we have examined (Go model 48-, 64-, and 80-mers, data not shown). These multiple transition state classes (and their corresponding intermediate phases) reflect the existence of parallel folding pathways, since our polymer may fold through any one of these transition state classes in a given folding event. Parallel unfolding pathways, with structually dissimilar, on-pathway intermediates have recently been observed in barstar [39].

The contribution of a given class to the transition state ensemble (and the importance of its corresponding intermediate to the folding pathway) can be modified by energetically stabilizing its contacts relative to other classes (V.S.P., Nik Putnam, and D.S.R., unpublished results). The existence of several independent pathways for a given structure, whose statistical weight can be tuned by altered connectivity or energetic stability, is consistent with experiments showing that cyclic permutants of the α-spectrin SH3 domain, while sharing the same native fold, can have different folding transition states [40].

A "classical" folding pathway

We have characterized the folding pathway of a protein-like heteropolymer on a three-dimensional lattice by directly simulating a collection of folding events. This polymer is "protein-like" in the sense that it rapidly folds to a reproducible, native state from any of a vast number of unfolded conformations. By determining the folding mechanism of such a simplified model, we can study in detail the manner in which the Levinthal entropy of the unfolded state is lost as the polymer folds.

The particular polymer we have studied folds *via* well-defined, partially folded, "on-pathway" intermediates, with a transition state ensemble described by specific core residues augmented by a limited range of optional contacts. This is consistent with a "classical" pathway, [35, 42, 43] which envisions protein folding as an intramolecular chemical reaction that proceeds from the unfolded state to the native state through a sequence of transiently populated intermediates. The species along this pathway are to be viewed as thermodynamically metastable phases of the polymer, separated by cooperative transitions.

METHODS.

Go model. We adopt the Go model [2] for protein-like heteropolymers, which has been widely used for both lattice [12, 19] and off-lattice [44] simulations. In this model, the "energy" of each polymer conformation is taken to be proportional to the number of (nearest-neighbor) native contacts it possesses, and non-native contacts incur no energetic cost or gain: $E(C) = -\epsilon Q$, where ϵ is the energy per native contact. It is convenient to measure temperature in units of ϵ/R. Note that the "energy" of a lattice polymer can be thought of as an "internal free energy" [12] that models the intra-polymer and polymer-solvent interactions as well as the solvent (and effective side-chain) entropy for the backbone conformation of interest.

By construction, the native state is the lowest energy conformation of the polymer. The Go model is known to exhibit a large energy gap

between native and other unrelated conformations [5], and to fold rapidly to its native state [35]. This model embodies the principle of "minimal frustration" [2, 8, 12], in the sense that the driving force for folding is the formation of native contacts, and there are no energetic barriers to achieving the folded state. For this reason the Go model cannot address issues pertaining to misfolded intermediates that are stabilized by non-native contacts. (A Go polymer could, in principle, exhibit an entangled misfolded state, in which some native contacts are formed in a context of a non-native fold. We have not observed this behavior in our simulations.) The fact that misfolded intermediates do not occur in the Go model, however, neither requires nor precludes it from exhibiting on-pathway intermediates, which is the issue we address here.

Native State. A specific compact conformation on a cubic lattice is selected as the "native" conformation; the folding events discussed here pertain to the native 48-mer conformation shown in the last frame of Fig. 1. (This structure was originally used in ref. [45].) We have not found any qualitative difference between the nature of the folding pathways for other choices of the native conformation (data not shown). The number of contacts that a given conformation shares with the native state is designated $Q \equiv \sum C_{ij} C_{ij}^{\mathrm{N}}$, where C_{ij} is the contact map of the conformation: $C_{ij} = 1$ if residues i and j are nearest neighbors in space (but not consecutive along the chain) and 0 otherwise. C_{ij}^{N} is the contact map of the native state.

Dynamics. The dynamics of the polymer chain are modeled by a Metropolis Monte Carlo process at temperature T [48]. Local movements of the polymer are considered, and always accepted if the new conformation has lower energy, but only accepted with probability $e^{-(E_{\mathrm{new}} - E_{\mathrm{old}})/RT}$ if the energy is increased by the move. Moves include local corner and crankshaft moves as discussed in ref. [6]. Monte Carlo dynamics includes both a bias towards lower energies as well as random thermal forces that model energy transfer to and from the solvent, which is assumed to be a heat bath in equilibrium at temperature T. The advantages and potential pitfalls of using such a scheme to model polymer dynamics are reviewed in ref. [46]. Simulations are performed on a Cray T3E at NERSC.

Heat capacity. From elementary statistical mechanics [47], the heat capacity $T\partial S/\partial T$ of a system in equilibrium is proportional to its time-averaged energy variance $\langle E^2 \rangle - \langle E \rangle^2$. Since in our model energy is proportional to Q, we define $c(t) \equiv \langle Q^2 \rangle - \langle Q \rangle^2$. We compute a time-dependent heat capacity $c(t)$ by calculating this variance over conformations sampled within $t \pm 1.5 \times 10^4$ steps. Since in our calculation the "energy" is actually a potential of mean force, heat capacity peaks reflect the release of *polymeric* entropy into the environment (*i.e.*, solvent), and are analogous to "latent heat" spikes found at phase transitions.

Folding probability. The folding probability p_{fold} of a given conformation $\mathcal{C}$ is determined by a set of additional simulations that are all started from conformation $\mathcal{C}$ and allowed to evolve in time. Each simulation is stopped when it reaches either a nearly folded or unfolded conformation. (For these purposes, a conformation is deemed folded if $Q/Q_{\max} \geq 90\%$ and unfolded if $Q/Q_{\max} \leq 20\%$. We typically use 400 runs to determine p_{fold}, yielding a statistical error of 5%.) The folding probability of conformation $\mathcal{C}$ is defined as the fraction of these simulations that fold before they unfold. p_{fold} is highly conformation-dependent. Fig. 1 reports the average p_{fold} for 500 conformations selected from a time interval of 10^4 steps.

Free energy. The total free energy $G(Q, Q_I)$ is computed by Monte Carlo sampling using a long equilibrium run (10^9 steps) during which many spontaneous folding and unfolding events occur. Free energy is measured by $G(Q, Q_I) = -RT \ln Z(Q, Q_I)$, where $Z(Q, Q_I)$ is the number of conformations sampled in the long Monte Carlo run that have the specified number of native and intermediate contacts. Regions of the (Q, Q_I) plane with low probability can be evaluated using the standard technique of umbrella sampling [48].

Acknowledgements. We thank Arup Chakraborty, David Chandler, Aaron Chamberlain, Alexander Grosberg, Tanya Raschke, and Susan Marqusee for useful discussions. This work was supported by the Miller Insitute for Basic Research in Science, National Science Foundation grant DMR-91-57414 and Lawrence Berkeley National Laboratory grant LDRD-3669-57. We acknowledge the use of the Cray T3E at the National Energy Research Scientific Computing Center.

*****

1. Shortle, D., Wang, Y., Gillespie, J.R. & Wrabl, J.O. Protein folding for realists: a timeless phenomenon. *Protein Science* **5**, 991-1000 (1996).
2. Ueda, Y., Taketomi H. & Go N. Studies on protein folding, unfolding, and fluctuations by computer simulation I. *Int. J. Peptide. Res.* **7**, 445-459 (1975).
3. Lau, K.F. & Dill, K.A. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* **22**, 3986-3997 (1989).
4. Skolnick, J. & Kolinski, A. Dynamic Monte Carlo simulations of globular protein folding/unfolding pathways. I. Six-member, Greek key beta-barrel proteins. *J. Mol. Biol.* 212, 787-817 (1990).
5. Sali A., Shakhnovich E.I. & Karplus M. How does a protein fold? *Nature* **369**, 248-51 (1994).
6. Pande, V.S., Grosberg, A. Yu, & Tanaka, T. Folding thermodynamics and kinetics of imprinted renaturable heteropolymers. *J. Chem. Phys.* **101**, 8246-8257 (1994).
7. Abkevich V.I., Gutin A.M. & Shakhnovich E.I. Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* **33**, 10026-10036 (1994).
8. Onuchic J.N., Socci N.D., Luthey-Schulten Z. & Wolynes P.G. Protein folding funnels: the nature of the transition state ensemble. *Folding & Design* **1**, 441-50 (1996).
9. Chan, H.S. & Dill, K.A. Protein folding in the landscape perspective: Chevron plots and non-Arrhenius kinetics. *Proteins Struct. Func. Gen.* **29**, to appear (1997).
10. Baldwin, R.L. The nature of protein folding pathways: the classical versus the new view. *J. Biomol. NMR* **5**, 103-9 (1995).
11. Wolynes, P.G., Onuchic, J.N. & Thirumalai, D. Navigating the folding routes. *Science* **267**, 1619-20 (1995).
12. Dill, K.A. & Chan, H.S. From Levinthal to pathways to funnels. *Nature Struct. Biol.* **4**:10-9 (1997).
13. Fersht A.R. Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proc. Nat. Acad. Sci. USA* **92**, 10869-10873 (1995).
14. Fersht A.R. Nucleation mechanisms in protein folding. *Curr. Op. Struct. Biol.* 7, 3-9 (1997).
15. Dill, K.A. & Shortle, D. Denatured states of proteins. *Annu. Rev. Biochem.* **60**, 795-825 (1991).
16. Shortle, D. The denatured state (the other half of the folding equation) and its role in protein stability. *Faseb Journal* **1**, 27-34 (1996).
17. Zhang, O. & Forman-Kay, J.D. NMR studies of unfolded states of an SH3 domain in aqueous solution and denaturing conditions. *Biochemistry*, **36**, 3959-70 (1997).
18. Fersht, A.R. Characterizing transition states in protein folding: an essential step in the puzzle. *Curr. Op. Struct. Biol.* **5**, 79-84 (1995).
19. Du, R., Pande. V.S., Grosberg, A.Yu, Tanaka, T. & Shakhnovich, E.I. On the transition coordinate for protein folding. *J. Chem. Phys.* **108**, 334-315 (1998).

20. Creighton, T.E., Darby, N.J. & Kemmink, J. The roles of partly folded intermediates in protein folding. *FASEB Journal* **10**, 110-118 (1996).
21. Abkevich, V.I., Gutin, A.M., & Shakhnovich, E.I. Free energy landscape for protein folding kinetics: Intermediates, traps, and multiple pathways in theory and lattice model simulations. *J. Chem. Phys.* **101**, 6052-6063 (1994).
22. Mirny, L.A. Abkevich, V., & Shakhnovich, E. I. Universality and diversity of the protein folding scenarios: a comprehensive analysis with the aid of a lattice model. *Fold. Des.* **1**, 103-116 (1996).
23. Abkevich, V.I., Gutin, A.M., & Shakhnovich, E.I. Domains in folding of model proteins. *Protein Science* **4**, 1167-1177 (1995).
24. Guo, Z., & Thirumalai, D. Kinetics and thermodynamics of folding of a de novo designed four-helix bundle protein. *Journal of Molecular Biology* **263**, 323-343 (1996).
25. Bai, Y., Sosnick, T.R., Mayne, L. & Englander, S.W. Protein folding intermediates: native state hydrogen exchange. *Science* **269**, 192-197 (1995).
26. Chamberlain, A.K., Handel T.M. & Marqusee, S. Detection of rare partially folded molecules in equilibrium with the native conformation of RNaseH. *Nature Struct. Biol.* **3**, 782-787 (1996).
27. Raschke, T.M. & Marqusee, S. The kinetic folding intermediate of ribonuclease H resembles the acid molten globule and partially unfolded molecules detected under native conditions. *Nature Struct. Biol.* 4, 298-304 (1997).
28. Kuwajima, K. The molten globule as a clue for understanding the folding and cooperativity of globular protein structure. *Proteins* **6**, 87-103 (1989).
29. Dobson, C.M. Protein folding: solid evidence for molten globules. *Curr. Biol.* **4**, 636-640 (1994).
30. Ptitsyn, O.B. Structures of folding intermediates. *Curr. Op. Struct. Biol.* **5**, 74-78 (1995).
31. Pande, V.S. & Rokhsar, D.S. Is the molten globule a third phase of proteins? *Proc. Nat. Acad. Sci., USA* **95**, 1490-1494 (1998).
32. Ptitsyn, O.B., Pain, R.H., Semisotnov, G.V., Zerovnik, E. & Razgulyaev, O.I. Evidence for a molten globule intermediate early in the kinetic folding pathway of apomyoglobin. *Febs. Lett.* **262**, 20-24 (1990).
33. Jennings, P.A. & Wright, P.E. Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin. *Science* **262**, 892-896 (1993).
34. Arai, M. & Kuwajima, K. Rapid formation of a molten globule intermediate in refolding of α-lactalbumin. *Folding & Design* **1**, 275-287 (1996).
35. Pande V.S., Grosberg A.Yu, Tanaka, T. & Rokhsar, D.S. Pathways for protein folding: Is a "new view" needed? *Curr. Op. Struct. Biol.* **8**, 68-79 (1998).
36. Otzen, D.E., Itzhaki, L.S., elMasry, N.F., Jackson, S.E. & Fersht, A.R. Structure of the transition state for the folding/unfolding of barley chymotrypsin inhibitor 2 and its implications for mechanisms of protein folding. *Proc. Nat. Acad. Sci., USA* **91**, 10422-10425 (1994).
37. Huang, G.S, Oas, T.G. Sub-millisecond folding of monomeric lambda repressor. *Proc Nat Acad Sci, USA* **92**, 6878-6882 (1995).
38. Schindler, T., Herrler, M., Marahiel, M.A., & Schmid, F.X. Extremely rapid protein folding in the absence of intermediates. *Nat. Struct. Biol.* **2**, 663-673 (1995).
39. Zaidi, F.N., Nath, U., & Udgaonkar, J.B. Multiple intermediates and transition states during protein unfolding. *Nat. Struct. Biol.* **4**, 1016-1024 (1997).
40. Viguera, A.R., Serrano, L. & Wilmanns, M. Different folding transition states may result in the same native structure. *Nature Struct. Biol.* **3**, 874-880 (1996).
41. Heidary, D.K., Gross, L.A., Roy, M. & Jennings, P.A. Evidence for an obligatory intermediate in the folding of interleukin-1β. *Nature Struct. Biol.* **4**, 725-31 (1997).
42. Kim, P.S. & Baldwin, R.L. Intermediates in the folding reactions of small proteins. *Annu. Rev. Biochem.* **59**, 631-660 (1990).
43. Matthews, C.R. Pathways of protein folding. *Annu. Rev. Biochem.* **62**, 653-683 (1993).
44. Nymeyer, H., Garcia, A.E., Onuchic, J.N. Folding Funnels and Frustration in Off-Lattice Minimalist Protein Landscapes. *preprint* (1998).
45. Shakhnovich, E., Abkevich, V., Ptitsyn, O. Conserved residues and the mechanism of protein folding. *Nature* **379**, 96-98 (1996).
46. Shakhnovich, E.I. Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29-40 (1997).
47. Chandler, D. *Introduction to Modern Statistical Mechanics.* (Oxford University Press, 1987).
48. Frenkel, D. & Smit, B. *Understanding Molecular Simulations.* (Academic Press, 1996).